

Articulator-Level Cross-Lingual Transfer in Sign Language Translation

Ali Elaswad*

The American University in Cairo
alyelaswad@aucegypt.edu

Omar Mokhtar Gaber Saqr*

The American University in Cairo
omar_saqr@aucegypt.edu

Abstract

Cross-lingual transfer is standard in low-resource sign language translation, but it is unclear what the transferred model provides. We fine-tune a pose-based model from American Sign Language (ASL) to Indian Sign Language (ISL) and decompose the transfer across its encoder, decoder, and articulator streams. The matched encoder-decoder pair reaches 3.61 ± 0.12 BLEU-4 over three seeds, while either component alone performs close to training from scratch. Content-word recall shows a sharp frequency effect: frequent content improves with training and data, whereas rare and unseen content stays near zero, even after twenty epochs and 49% more data. This is not a decoder artefact: decoded blind, the same checkpoint reproduces no frequency gradient, and frequency still orders recall within strata of contextual predictability. Comparing ASL and Chinese Sign Language sources with the output language held fixed, a 22% encoder advantage survives across seeds, and multi-seed weight analysis localises it to the hand, spatial, and body components; an apparent Chinese advantage on face and temporal does not survive a second seed. A second ASL corpus of a different domain reproduces the effect. Transfer supplies a coupled encoder-decoder starting point, but does not by itself solve the long tail of target-language content.

1 Introduction

Sign language translation has drawn increasing research interest, but progress remains concentrated on a small number of languages with large annotated corpora. For low-resource languages such as Indian Sign Language (ISL), a common strategy is

cross-lingual transfer: pretrain on a higher-resource sign language and fine-tune on the target (Artiaga et al., 2024; Dal Bianco et al., 2025). Yet papers typically report only the resulting aggregate score, leaving open what the transferred model actually provides.

We study this question using a pose-based encoder-decoder model fine-tuned from American Sign Language (ASL) to ISL. We first decompose the transferred checkpoint by separating its sign encoder and language decoder. The matched pair reaches 3.61 ± 0.12 BLEU-4 over three seeds, while either component alone performs close to training from scratch. We then examine what the resulting model can actually translate. Content-word recall follows training frequency closely: frequent content improves with training and additional data, while rare and unseen content remains near zero even after twenty epochs and a 49% increase in training data. Two controls establish that this is a property of what the model reads rather than of its decoder: with the visual input removed the same checkpoint shows no frequency gradient, and frequency continues to order recall within strata of contextual predictability.

We additionally compare ASL and Chinese Sign Language (CSL) as source languages while holding the output language fixed. Most of the apparent gap between full checkpoints turns out to be the decoder’s output language, leaving a 22% encoder advantage whose per-arm ranges do not overlap across seeds. Running the weight analysis over several seeds separates what survives from what does not: the ASL encoder starts measurably closer to ISL on the hand, spatial, and body components, while an apparent CSL advantage on the face and temporal components lies inside run-to-run spread. Because the ASL and CSL checkpoints differ in pretraining domain as well as language, we treat the surviving pattern as suggestive; a second ASL corpus of a different domain reproduces it, which

* These authors contributed equally to this work.

The iSign corpus (Joshi et al., 2024) and the Uni-Sign checkpoints (Li et al., 2025) we build on are public; Appendix A gives the source checkpoints, training configuration, per-condition initialisation, and evaluation protocol behind every number here.

weakens a purely domain-based account without settling the question. Figure 1 further shows that the transferred model depends strongly on the body stream: removing it causes a larger degradation than removing either hand, while the hands alone provide almost no translation signal.

Our contributions are:

- We show that transfer from ASL to ISL requires the matched pretrained encoder-decoder pair; neither component alone recovers the transfer gain.
- We show that content-word recall is strongly tied to training frequency, with rare and unseen content remaining resistant to additional training and data.
- We isolate source-language effects from decoder output language and find an articulator-specific pattern, reported over multiple seeds so that the components which survive run-to-run variation are separated from those which do not, and whose interpretation remains limited by a source-domain confound.
- We audit our own evaluation: a corrected clip-to-video mapping exposing residual split leakage and its measured effect on every score, and two controls separating the vocabulary effect from the decoder’s language model.

2 Related Work

Sign language translation has drawn growing attention as a way to bridge deaf and hearing communities (Camgoz et al., 2020; Joshi et al., 2024). Early systems relied on gloss annotation as an intermediate supervision signal, but glosses are costly to produce and unavailable for most sign languages, which has pushed recent work toward gloss-free translation. A common recipe pairs a visual encoder over the signing video, or its extracted pose keypoints, with a large pretrained language model as decoder (Zhou et al., 2023; Wong et al., 2024; Li et al., 2025), a design that trades some appearance detail for efficiency and signer invariance when keypoints replace raw frames (Dal Bianco et al., 2025).

Two properties of the languages bear on what follows. Sign languages mark interrogatives and negation non-manually, through brow position held across a clause rather than a lexical item (Zeshan,

2004), which makes the face stream a meaningful unit to ablate. Indian Sign Language is typologically independent of both American and Chinese Sign Language (Zeshan, 2003), so neither source is a close relative. The corpora also differ in domain — How2Sign is instructional monologue (Duarte et al., 2021), OpenASL interpreted news (Shi et al., 2022), CSL-Daily scripted conversation (Zhou et al., 2021) — a difference we use as a partial control in Section 4.4.

Because sign language data is scarce, a substantial line of work looks outside the target language for supervision. Scaling the pretraining corpus, model size, and number of source languages together improves translation broadly, though quality remains far from usable (Zhang et al., 2024). Multilingual pretraining followed by brief target-language fine-tuning reaches 2.71 BLEU-4 from a 2.61 monolingual baseline on ISLTranslate (Joshi et al., 2023), a different ISL corpus, and is the closest prior *transfer* result to ours (Dal Bianco et al., 2025). A complementary approach stitches poses for words drawn from recognition datasets into synthetic training pairs, raising iSign — the corpus we use — from 0.55 to 3.43 (Joshi et al., 2025); because it can only recombine signs it already has isolated examples for, it bounds what lexicon-internal augmentation can buy. Transfer between sign languages is established for isolated-sign recognition (Artiaga et al., 2024), but translation, with its encoder-decoder structure and open output space, raises the further question of which components carry the benefit. Pose is not the only route: video features augmented with optical flow reach 8.58 BLEU-4 on iSign (Chowdhury and Sanyal, 2025), well above any pose-only system including ours. We study what transfer contributes within a pose-only architecture, not what the corpus admits overall.

Two recent studies bear on our findings without addressing them directly. A self-supervised model over separate hand, face, and body streams reports a masking ablation on ASL consistent with ours: hands alone reach 0.2 BLEU-4, comparable to blind decoding, against 2.4 for all streams (Gueu-wou et al., 2025). At a stronger operating point on British Sign Language, remaining failure modes include an inability to distinguish questions from statements, missed negations, and difficulty with fingerspelling (Jang et al., 2025) — all of which we observe independently.

Closest to our concerns about evaluation, signer-

dependent splits are shown to inflate reported scores, with BLEU-4 on one benchmark falling by more than half under signer-independent evaluation (Artiaga et al., 2025). We split by source video for this reason; Section 4 audits how well our implementation of that rule held and what its residual failure was worth. Signer identity is not distributed with the corpus, so signer non-overlap cannot be checked directly. Our benchmark anticipates a related concern, noting that names are fingerspelled and then assigned a short sign, and that clip segmentation often separates a name’s introduction from its later use (Joshi et al., 2024).

3 Methodology

3.1 Model

We use Uni-Sign (Li et al., 2025) in its pose-only configuration. Four ST-GCN streams (Yan et al., 2018), one each for body, left hand, right hand, and face, read pose keypoints and produce pooled per-stream representations. These are concatenated and projected into an mT5-base decoder (Xue et al., 2021), which we refer to throughout as the sign encoder and the language decoder respectively. The encoder reads 69 of the 133 COCO-WholeBody keypoints (Jin et al., 2020): 9 body, 21 per hand, and 18 face. Uni-Sign ties the left- and right-hand modules, so the two hands share one set of weights; we report them jointly as *hands* throughout. The two halves are architecturally separable: either can be loaded from a pretrained checkpoint while the other is left at its own pretrained initialisation, or replaced entirely with a checkpoint pretrained on a different source language. This separability is what makes the decomposition and source-language experiments in Section 4 possible without any change to the architecture itself. Where we test the pretrained encoder in isolation, we keep it frozen; this probes the pretrained representation directly rather than testing whether a trainable encoder-only configuration would also underperform.

Source checkpoints. We do not pretrain: all three source checkpoints are the public pose-only translation releases accompanying Li et al. (2025), which share a common stage-1 pose pretraining on CSL-News and differ only in the translation corpus each was subsequently fine-tuned on. We use `how2sign`, `openasl`, and `csl_daily`, trained respectively on How2Sign (Duarte et al., 2021), OpenASL (Shi et al., 2022), and CSL-Daily (Zhou et al., 2021). Because stage 1 is shared,

a comparison between two of these checkpoints isolates the stage-2 translation corpus, which is what varies in sign language and in domain; it does not isolate language from domain, and Section 4.4 returns to that. Appendix A lists file identifiers, parameter counts, and exactly which tensors each condition loads, freezes, and trains.

3.2 Isolating the Source Language

A full checkpoint pretrained on one sign language differs from a full checkpoint pretrained on another in two ways at once: the sign language the encoder was trained to read, and the spoken language the decoder was trained to write. Comparing two full checkpoints therefore conflates both effects, and cannot tell us which one is responsible for an observed gap.

We separate them by constructing a hybrid model: the sign encoder from one source-language checkpoint, paired with the language decoder from another. Concretely, we pair a Chinese Sign Language (CSL) encoder with the English-writing decoder from the ASL checkpoint. This hybrid isolates the encoder’s source-language effect, since the decoder’s output language is now held constant across every comparison. We separately confirmed that the ordering between source languages was unchanged when one arm was rerun under the other’s training configuration, which rules out incidental training differences as the explanation for the ordering; we did not, however, control for the fact that the two source checkpoints’ pretraining data differ in domain as well as language, which we return to in Section 4.4.

3.3 Locating the Advantage Across Articulators

Where a source-language advantage exists, we ask where in the signer’s body it lives, using two methods that draw on different signals. The first measures how much each source encoder’s weights change during ISL fine-tuning, separately for each of five components: hands, spatial configuration, body, face, and temporal movement. We define relative weight change for component c as

$$\Delta_c = \frac{\|\theta_c^{\text{fine-tuned}} - \theta_c^{\text{pretrained}}\|}{\|\theta_c^{\text{pretrained}}\|} \times 100\%, \quad (1)$$

where $\theta_c^{\text{pretrained}}$ is the component’s weights immediately after loading the source-language checkpoint and $\theta_c^{\text{fine-tuned}}$ is the same component’s

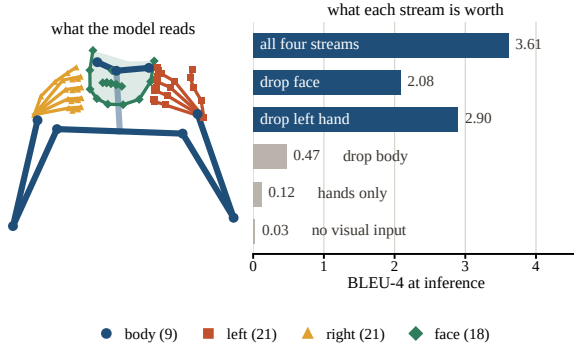


Figure 1: Left: keypoints supplied to each pose stream for a test clip. Right: BLEU-4 when streams are zeroed at inference. Removing the body stream causes the largest single-stream degradation, while the hands alone provide almost no translation signal.

weights after ISL fine-tuning completes. A smaller Δ_c means the source encoder’s representation for component c needed less adjustment to fit ISL. Integer buffers such as BatchNorm step counters are excluded; squared in a norm, they would otherwise dominate every component.

The five components are *two overlapping partitions* of the same encoder, not five disjoint parts, and they do not sum. Hands, body, and face partition it by stream (74 tensors each); spatial and temporal partition it by layer type within the ST-GCN blocks (132 and 156 tensors), cutting across all four streams. The temporal component shows the largest relative change in every run, partly because its pretrained weights have the smallest norm, so an equal displacement registers as a larger ratio. Δ_c is therefore comparable between two sources *within* a component, not between components, and we treat it as directional rather than calibrated evidence: there is no control arm against which “small” is defined. We report it over two ASL and two CSL runs in one training configuration, and call a component separating only when the per-arm ranges do not overlap.

That second method measures behavioural recall directly: for content tied to a specific articulator, hand-carried proper nouns and numerals, or face-marked interrogatives, we measure how often the fine-tuned model reproduces that content correctly, with the decoder held fixed as described above. We define content-word recall as

$$R_{\text{word}} = \frac{1}{|D|} \sum_{d \in D} \frac{|\hat{w}_d \cap w_d|}{|w_d|}, \quad (2)$$

where D is the evaluation set of clips, w_d is the set

of target content words for clip d , and $\hat{w}_d$ is the set of content words the model actually produces for that clip. Question recall, reported in Sections 4.4 and 4.7, is a distinct quantity: the fraction of interrogative references rendered as questions in the output, and is not computed by Equation 2.

Both Δ_c and R_{word} are computed from the same underlying fine-tuning runs, so agreement between them is evidence that a pattern is not an artefact of one measurement method, but it is not evidence that the pattern is independent of confounds shared by both runs, including the pretraining domain confound noted above.

Finally, because the model’s four pose streams are concatenated rather than fused earlier in the network, any subset of them can be zeroed at inference without retraining. We use this to run an articulator masking ablation directly on the trained model, which gives a third, purely behavioural view of which streams the model actually needs, independent of any source-language comparison.

4 Experiments

4.1 Experimental Setup

We use a verified-representative subsample of the iSign pose corpus, 23,921 of its 127,237 clips: perdecile counts range from 2332 to 2457 against an expected 2392, and mean reference length matches the corpus to within 0.04 words, 10.76 against 10.72. iSign defines no canonical split; we follow the dataset authors’ recommendation to split by source video, obtaining 19,093 training, 2,396 development, and 2,432 test clips. Published results on this benchmark use self-constructed splits, so cross-paper comparisons throughout this section are indicative rather than exact.

Split audit. Our first clip-to-video rule was wrong, fragmenting single videos whose identifiers contain hyphens (Appendix B). Corrected, the splits hold 3,198, 411, and 413 source videos, and 22 videos cross the train–test boundary, contributing 0.90% of test clips. Removing them *raises* BLEU-4 by +0.02 to +0.03 on all three ASL seeds, since those clips are harder than average: the leakage is real but inflates nothing. Signer identity is not released with iSign, so signer-level independence cannot be checked directly.

Inputs are converted from MediaPipe Holistic (Lugaresi et al., 2019) to the COCO-WholeBody-133 layout the model expects (Jin et al., 2020). Every condition trains for 20 epochs with AdamW

(Loshchilov and Hutter, 2019) under a one-cycle schedule at peak learning rate 3×10^{-4} , effective batch 8, and bf16 autocast, on a single 16 GB or 24 GB consumer GPU. We take the final epoch rather than selecting on development score, so no condition benefits from checkpoint selection the others did not get; Appendix A gives the full configuration. Evaluation uses beam search with 4 beams and a 100-token limit, matching the reference implementation, and we report BLEU-4 (Papineni et al., 2002) and chrF (Popović, 2015) through sacrebleu (Post, 2018).

Seeds, and which set numbers come from. Every number below is a full test-set score, and conditions run at up to three seeds; we report the mean, the population standard deviation, and the individual runs. With so few runs, that spread describes the runs we have rather than a sampling distribution, so our claims turn on whether two arms’ ranges overlap. Arms are compared only within a training configuration, since micro-batch size changes padding (Appendix A). Inference-time analyses needing one fixed model (Sections 4.5–4.6) use the seed-13 checkpoint, 3.58 BLEU-4 and 21.85 chrF standalone; bf16 decoding is not bit-deterministic, so it varies by up to ± 0.03 BLEU-4 between passes and each table carries its own control.

We validated the evaluation harness independently: the released ASL checkpoint scores 9.59 BLEU-4 and 30.28 chrF on How2Sign through our pipeline, matching its published figure, so the low absolute scores that follow reflect the difficulty of the ISL task rather than a fault in the harness.

4.2 Transfer Requires Both Components

Table 1 decomposes the transferred checkpoint on the test set. Every partial transfer we can construct, whether half the network or a single pose stream, lands between 0.31 and 0.76 BLEU-4, a band that straddles training from scratch rather than sitting above it, while the matched pair reaches 3.61 ± 0.12 : roughly five times the best of them. Under this decomposition, the transferred benefit is not recovered by either component in isolation, but by the matched encoder-decoder pair. This has a direct practical consequence: replacing either half after transfer, for instance swapping in a stronger decoder, discards this relationship even when the replacement is individually stronger on its own.

One caveat: the frozen-encoder arm has by far the widest spread, 1.12 against 0.41 and 0.47, so

Transferred	BLEU-4	Runs
Encoder + decoder	3.61 ± 0.12	3.58, 3.77, 3.47
Encoder only (frozen)	0.67 ± 0.32	1.12, 0.41, 0.47
Decoder only	0.76	0.76
Hand streams only	0.61 ± 0.08	0.70, 0.53
Body stream only	0.58	0.58
Face stream only	0.31	0.31
Neither (scratch)	0.455 ± 0.001	0.454, 0.456

Table 1: Test-set decomposition of an ASL→ISL transfer. No partial transfer materially exceeds scratch; only the matched pair does. Spread is population standard deviation; the scratch row is given to three decimals because its two runs differ by 0.002. chrF in Appendix A. For reference, Joshi et al. (2025) report 3.43 BLEU-4 on iSign, the corpus we use, and Dal Bianco et al. (2025) 2.71 via multilingual transfer on ISLTranslate.

its mean of 0.67 rests on one high run and is not a tier above the other partial transfers. What is robust is the gap between *any* partial transfer and the matched pair, an order of magnitude beyond the largest within-arm spread (Appendix A notes the configurations).

4.3 Source Language, Isolated from Output Language

Table 2 reports the intact CSL checkpoint alongside the hybrid configurations described in Section 3. Comparing full checkpoints within one configuration, ASL reaches 3.61 and CSL 0.59: an apparent gap of 3.02 BLEU-4, almost all of it the decoder’s output language. With the decoder held fixed and both arms in the same configuration, the gap is 0.68 (3.75 against 3.07) — a 22% encoder advantage, not a fivefold one. That 22% is not a single-run artefact: the two arms’ ranges, 3.67–3.82 against 3.06–3.09, do not overlap.

Both checkpoints share stage-1 pretraining and differ in the sign language *and* the domain of the stage-2 corpus. A partial control is available: OpenASL is also ASL but is interpreted news rather than instructional monologue, differing from How2Sign in domain about as much as CSL-Daily does. Its encoder reaches 3.54 with the same English decoder, close to How2Sign’s 3.61 in that configuration and well clear of CSL’s 3.07, while the configuration difference is worth only +0.14. Two ASL corpora of different domains behave alike and both beat CSL, which weakens a purely domain-based account without eliminating it: both remain more hand-forward than scripted conversation.

Encoder	Dec.	<i>n</i>	BLEU-4	chrF
<i>matched configuration, same machine</i>				
How2Sign (ASL)	En	2	3.75 ± 0.08	21.82
CSL-Daily (CSL)	En	2	3.07 ± 0.01	20.39
<i>second configuration</i>				
How2Sign (ASL)	En	3	3.61 ± 0.12	21.80
OpenASL (ASL)	En	1	3.54	21.82
CSL-Daily (CSL)	Zh	1	0.59	15.61

Table 2: Test scores separating source sign language from output language. Most of the naive ASL–CSL gap is the decoder; a 22% encoder advantage survives with the decoder held fixed, with non-overlapping per-arm ranges. OpenASL is a second ASL source of a different domain.

4.4 Where the Advantage Lives

We report this analysis as a suggestive pattern, for the reasons noted in Section 3. Table 3 and Figure 2 report weight change by component over two ASL and two CSL runs, all in one training configuration.

ASL starts closer to ISL on hands, spatial configuration, and body, with ranges that do not overlap. Face and temporal do not separate: face spans 415–492 across ASL runs against 365–515 across CSL, and temporal 264–293 against 281–326. A single run per source had shown CSL closer on both, which invited a typological reading of that split; it was an artefact of that run’s ASL face value, 718.6, and we withdraw the claim. A third ASL run in the other configuration leaves which components separate unchanged.

Behavioural recall, computed as in Equation 2 with the decoder held fixed and paired at the item level, shows the same pattern: ASL leads on proper nouns (20.7% against 18.2%, 95% CI [+1.4, +3.6]) and numerals (35.9% against 29.3%, [+3.9, +9.3]), content carried mostly by hand shape, while face-marked questions do not separate (36.9% against 31.0% on 84 clips, interval [−2.4, +14.9]). Two measurements on different signals agree on which components resolve.

One explanation is that ASL is linguistically closer to ISL in hand configuration. An equally consistent explanation is that our ASL source is instructional monologue, which tends to be lexically dense and hand-forward, while our CSL source is scripted conversation. The OpenASL arm of Section 4.3 favours the first without settling it, and we flag this as the main open question raised by our source-language experiments rather than resolving it here.

Component	ASL (<i>n</i> =2)	CSL (<i>n</i> =2)	Closer
Hands	142.1 ± 1.6	157.9 ± 3.2	ASL
Spatial	50.5 ± 2.3	58.2 ± 1.9	ASL
Body	174.5 ± 7.8	217.1 ± 21.1	ASL
Face	453.9 ± 38.5	440.0 ± 74.9	—
Temporal	278.4 ± 14.3	303.3 ± 22.8	—

Table 3: Relative weight change during ISL fine-tuning (Δ_c , Equation 1), mean and population standard deviation over two runs per source in one configuration. Smaller means the source’s pretrained representation needed less adjustment for ISL. “—” marks components whose per-arm ranges overlap, so the two sources do not separate. Hands, body, and face partition the encoder by stream; spatial (graph convolution over joints) and temporal (convolution over time) partition it by layer type, so the five rows are two overlapping views and do not sum.

Condition	BLEU-4	chrF	Date recall
Intact	3.59	21.80	65.8%
Shuffled poses	0.25	14.25	0.9%
Blind (all masked)	0.03	13.96	0.0%

Table 4: Grounding controls, test set, all decoded in one pass from the seed-13 checkpoint. Removing or scrambling the visual input destroys performance. The intact row is this pass’s own control and reads 3.59 against the 3.58 of standalone evaluation and the 3.61 of the masking pass in Appendix E; the spread is decode nondeterminism, not a different model.

4.5 Grounding

Before relying on any of the above, we check that the model is reading the signing rather than generating plausible English independent of it. Blind decoding zeroes every pose stream at inference. Shuffled decoding feeds each clip a different clip’s pose sequence while keeping its own reference, a stronger control since the input remains well-formed sign motion, just the wrong motion. We track date recall alongside BLEU, since dates cannot be guessed from a language prior and are exactly measurable.

Both controls collapse performance (Table 4): 0 of 111 dates survive blind decoding, and only 1 survives shuffling. Blind chrF, 13.96, falls below the length-matched random-sentence baseline of 15.14, meaning the model without visual input performs worse than random guessing calibrated to sentence length. The model’s output depends on reading the visual input, and that blind decode is reused in Section 4.8 as the language-prior control.

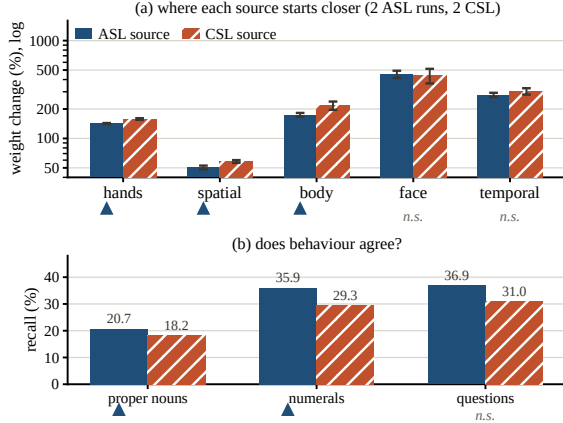


Figure 2: Where the ASL and CSL source encoders differ. (a) Relative weight change during ISL fine-tuning by component, log scale, two runs per source in one configuration, whiskers spanning their range. (b) Behavioural recall for content tied to each articulator, with the decoder held fixed. A triangle marks the source that wins; “n.s.” marks a comparison whose ranges overlap (a) or whose bootstrap interval spans zero (b). The two panels agree: hand-carried measures separate, face-marked ones do not.

4.6 Articulator Masking

Because the four pose streams are concatenated rather than fused, any subset can be zeroed at inference without retraining. Figure 1 illustrates the four input streams and the resulting BLEU-4 changes, while Appendix E reports all sixteen stream combinations with chrF and date recall. Every configuration lacking the body stream collapses.

The body stream is necessary but not sufficient. Above size one, a subset containing body beats every same-sized subset without it, and removing body from the full set costs $3.61 \rightarrow 0.47$, more than removing any other stream. Yet body *alone* reaches 0.05, below hands alone at 0.12 and indistinguishable from blind decoding; no single stream exceeds 0.10. What the model needs is the conjunction: body with one hand reaches 1.26 and recovers 55% of dates, while both hands without body reach 0.12 and recover 0.9%. Hand configuration appears interpretable only relative to where the hands are, which the architecture encodes by positioning the hand streams on body keypoints. The hands are also not interchangeable: removing the right costs $3.61 \rightarrow 1.09$ against 2.90 for the left.

4.7 Non-Manual Marking

Sign languages mark interrogatives non-manually, principally through eyebrow position held across

Seed	BLEU-4		Q recall	
	brow	jaw	brow	jaw
13	3.46	3.58	44.0%	33.3%
21	3.45	3.77	34.5%	39.3%
37	3.59	3.47	38.1%	42.9%
mean	3.50	3.61	38.9%	38.5%

Table 5: Eyebrow landmarks versus jaw contour, three seeds per arm. No measurable effect: both arms’ ranges overlap on both measures.

the clause (Zeshan, 2004). The face stream in this architecture samples jaw contour, inner mouth, and nose tip; mapped onto the standard 68-point face scheme (Sagonas et al., 2016), none of its 18 landmarks fall on the brows. Behaviour matches this gap: the seed-13 checkpoint renders 33.3% of interrogative references as questions, against 97.0% of statements kept as statements. (Section 4.4 quotes 36.9% for a different arm; 84 clips cannot resolve it.) Transfer lifts interrogative rendering (16.7% scratch, 25.0% frozen encoder, 33.3% matched pair) but it stays far behind the statement rate, and behind content grounding over the same three models, where outputs sharing no content word with the reference fall from 93.5% to 49.9%.

We tested the implied fix directly: refilling the jaw-contour slots with eyebrow landmarks, keeping the stream size fixed so pretrained weights still load, three seeds per arm with matched jaw controls (Table 5). It does not measurably help. Pooled over seeds, question recall is 38.9% with eyebrows against 38.5% without, a difference of 0.4 points ($z = +0.09$, $p = 0.93$), with translation quality unchanged. A single seed would have suggested otherwise: seed 13 alone gives eyebrows a 10.7-point advantage, pairing that arm’s best seed with the jaw arm’s worst. Within-arm spread, 34.5–44.0% for brow and 33.3–42.9% for jaw, dwarfs the 0.4-point pooled difference. Because the pretrained face weights were fit to jaw-contour statistics, we cannot separate a capacity limit from a structural mismatch, making this an observation rather than a controlled test. Both readings predict the same lesson: adding non-manual features without adapting the capacity devoted to them is unlikely to help.

4.8 Vocabulary

This is, independent of the source-language comparisons above, our most consequential finding. Content-word recall, computed as in Equation 2, tracks closely with how often each reference word

Behaviour	Reference	Model output
Anchors survive, context invented	The Queen of the United Kingdom – Queen Elizabeth passed away on the afternoon of 8th September 2022.	However in a devastating incident, a beautiful girl named Queen Elizabeth died on 8th September 2022.
Unseen name dropped	DWWF was founded in 2014 and has been working ever since.	It was founded in 2014 and has been working ever since.
Correct, BLEU-penalised	On 3rd July Rahul Gandhi has made his resignation official and final.	On 3rd July, Rahul Gandhi announced his retirement.
Polarity reversed	This is the first time he is visiting the state since the removal of Article 370 in 2019.	This is not the first time such a case was reported in 2019 and was not the first time such a case was reported.
Degenerate repetition	On 15th June, 20 Indian soldiers were martyred in a violent face-off with Chinese troops in Galwan Valley.	On 15th June, the Chinese army participated in a fight against the Chinese army.

Table 6: Test-set outputs from the ASL-transferred model, one per behaviour. Dates and frequent names survive, unseen names vanish without a trace, and the surrounding text stays fluent whether or not it is accurate.

Predict.	unseen	rare	mid	freq.	v. freq.
low	0.000	0.002	0.072	0.143	0.173
mid	0.002	0.010	0.079	0.172	0.232
high	—	0.056	0.163	0.246	0.271

Table 7: Content-word recall by training frequency within terciles of contextual predictability. Frequency orders recall inside every tercile, so the frequency effect is not reducible to formulaic context. “—” marks a cell with too few tokens to estimate (142 or fewer elsewhere in that row).

Training frequency	n	Intact	Blind
Unseen (0)	1,251	0.001	0.000
Rare (1–4)	1,907	0.009	0.000
Mid (5–49)	5,920	0.096	0.002
Frequent (50–499)	6,397	0.203	0.016
Very frequent (500+)	1,320	0.248	0.001

Table 8: Content-word recall by training frequency, with the same checkpoint decoded blind as a language-prior control. Words unseen in training are essentially never produced, and the blind column shows no frequency gradient: the effect belongs to what the model reads, not to its decoder.

appeared in training (Table 8), spanning more than two orders of magnitude between unseen and very frequent words.

Is this the decoder rather than the encoder? A frequent word is also likelier to sit in a formulaic context the decoder can pattern-match, so the curve could be a language-model effect. Two controls say it is not. Decoding the same checkpoint with every pose stream zeroed isolates the language prior, and that prior shows no frequency gradient: recall is ≤ 0.002 in four of five bands and peaks at 0.016, against 0.248 intact (Table 8, right column). Second, estimating each word’s contextual predictability from its left context under a trigram model fitted to the training references, frequency still orders recall inside every tercile and the unseen band stays at the floor in all of them (Table 7). A logistic fit on both standardised predictors puts log frequency at $\beta = +0.73$ against log predictability at $\beta = +0.25$.

Restricting to proper nouns, which must be fingerspelled and so give a cleaner test of reading than ordinary vocabulary, the pattern sharpens: 0 of 370 proper-noun tokens unseen in training are ever re-

produced, against 49.8% of those seen 50 times or more. We detect proper nouns orthographically, as capitalised tokens in non-sentence-initial position with possessives normalised, a rule with no dependence on a named-entity model; the same count run over the data-scale pair below gives 0 of 1,130 and 0 of 465. Figure 3 shows what this looks like in a single test clip: everything the model has seen in training survives, and the one word absent from training, a name, is replaced by a pronoun with no signal that anything is missing. Measured in aggregate, proper-noun recall of 0.216 looks reasonable; that aggregate is carried entirely by high-frequency names inferable from topic, and hides that recall on unseen names is exactly zero.

To interpret the headline score itself, we anchor it against trivial baselines that require no model at all (Appendix D). A random training sentence of matched length reaches chrF 15.14, close to what training from scratch achieves (15.29 over two seeds); the transferred model reaches 24 times the random baseline on BLEU-4, so the gain over these floors is real, even though the floors themselves sit

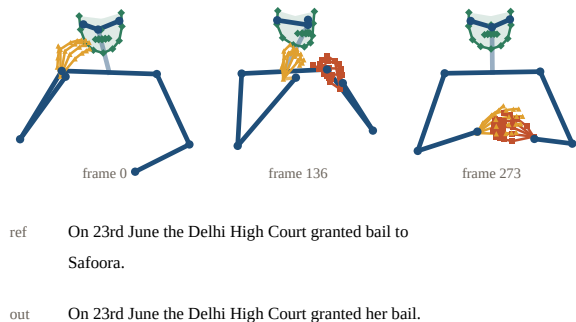


Figure 3: A test clip, evenly sampled, with its reference and the model’s output. Everything the model has seen in training survives: the date, the institution, the verb phrase. The one word absent from training, a name, is replaced by a pronoun, with nothing in the output marking that a name was there.

higher than an unanchored score would suggest.

The same pattern holds under scaling. A controlled comparison on a matched test set, one arm trained on 19,099 clips and the other on 28,512 (a separate corpus draw, Appendix A), shows recall rising at the frequent end while the unseen end stays at zero (Figure 4). Aggregate score does improve, from 3.31 ± 0.02 to 3.98 BLEU-4, a 21% relative gain from 49% more data, but recall on frequent words rises only from 0.214 to 0.230 and neither arm reproduces a single unseen proper noun. More training data behaves like more transfer: it improves the vocabulary the model already has some purchase on, not the vocabulary it has never seen. Our own headline result should be read in the same light. It uses 19,093 training clips, 15.0% of the 127,237-pair corpus (our full subsample is 18.8%). That makes the result data-efficient in aggregate score and not in any sense open-vocabulary, since both arms share the same floor.

We also checked whether pose estimation quality, rather than the model itself, explains this pattern. Stratifying test clips by hand-keypoint coverage rejects that explanation: the best coverage quartile reaches content recall of 0.152, against 0.108 in the worst, with correlation $r = +0.088$ across clips. Near-perfect hand tracking yields nearly identical failure, so improving keypoint extraction is not the lever it is often assumed to be.

4.9 Failure Modes

Table 6 shows five representative outputs, one per behaviour discussed above; Appendix C asks where the surrounding text comes from. Content the

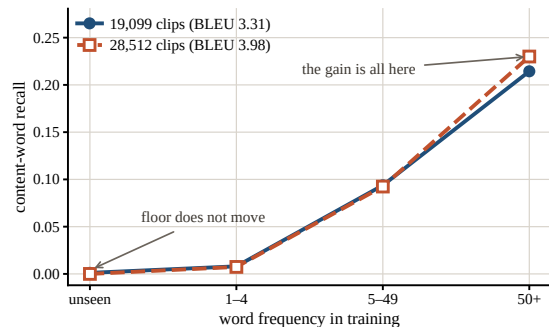


Figure 4: Content-word recall by training frequency for the controlled data-scale pair, evaluated on the same test set with frequency counted in each arm’s own training corpus. The smaller arm is the mean of two seeds. Half again as much data lifts the frequent end of the curve and leaves the unseen end pinned at zero.

model has seen in training survives fluently; content it has not seen is either dropped or replaced with a plausible substitute, and the resulting sentence gives no indication that anything is missing.

Two of these behaviours are worth separating from ordinary translation error. The third example is penalised by BLEU despite being an accurate paraphrase, which shows the metric is not uniformly too generous here. The fourth reverses the polarity of the source sentence while remaining fluent and confident, with nothing in the output to signal the reversal to a reader who cannot check the source. Across the test set, 9.9% of positive references are rendered negative, a rate high enough to matter for any deployment where a user cannot verify against the original signing.

5 Conclusion

Cross-lingual transfer improves low-resource sign language translation, and we report the best pose-only result we are aware of on Indian Sign Language from a 15% sample of iSign, though well short of video-based systems on the same corpus. Two findings follow from taking this result apart. First, the benefit comes from a transferred encoder-decoder relationship rather than from either component alone, a pattern robust across every comparison we ran. Second, and independent of any source-language comparison, the system’s recall tracks training frequency almost exactly and does not improve on rare or unseen content under any intervention we tested. The source-language advantage also splits unevenly across the body, holding on the hands but not the face.

Limitations

Our conditions run at up to three seeds, which is enough to show when an arm’s range is disjoint from another’s but not enough to estimate a sampling distribution. One checkpoint per source language remains a single sample of that language, so the 22% encoder advantage is a property of these two checkpoints before it is a property of ASL and CSL.

The ASL and CSL source checkpoints differ in domain as well as sign language. The OpenASL arm weakens a purely domain-based reading, since two ASL corpora of different domains behave alike, but both are still more hand-forward than scripted conversation, so we cannot fully separate the two explanations and present the articulator pattern as suggestive throughout.

Our splits contain residual source-video leakage: 22 videos cross the train–test boundary, contributing 0.90% of test clips. We measure its effect rather than assume it away, and removing the affected clips raises BLEU-4 slightly on every seed, so it is not inflating our results. Signer identity is not distributed with iSign, so signer-level independence cannot be verified directly and source-video separation is the strongest proxy available to us.

Our absolute operating point is low, 3.61 BLEU-4. The decompositions we report are therefore established in a regime where the model translates poorly, and encoder–decoder coupling, frequency-tracking recall, and body-stream dominance may behave differently in a system with more headroom. The vocabulary finding is the one we would most expect to generalise, since it concerns what is absent from the training distribution rather than what the model does with what is present, but we have not tested it elsewhere.

Our non-manual marking experiment is a negative result that admits two readings, as Section 4.7 notes: the substituted eyebrow landmarks reuse weights fitted to jaw-contour statistics, so a structural mismatch and a capacity limit predict the same null.

We study one architecture, one target language, and one transfer direction. The parameter asymmetry between visual and language components is characteristic of gloss-free pose-based sign language translation generally, but the specific face-landmark selection is one design choice among several. The non-manual marking finding in particular should be checked against other systems

before being treated as a general property of the architecture class.

Acknowledgments

Generative AI tools were used only for language editing and readability improvements. All research, experimental design, implementation, and results are the authors’ own work. We thank the anonymous reviewers, whose questions about reproducibility, split integrity, single-run design, and the interpretation of the vocabulary effect prompted the split audit in Appendix B, the multi-seed tables throughout Section 4, and the language-prior and predictability controls in Section 4.8. Three claims in the submitted version did not survive those checks and have been withdrawn.

References

- Keren Artiaga, Sabyasachi Kamila, Haithem Afli, Conor Lynch, and Mohammed Hasanuzzaman. 2025. [Re-thinking sign language translation: The impact of signer dependence on model evaluation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18379–18391.
- Keren Artiaga, Yang Li, Ercan Engin Kuruoglu, and Wai Kin Chan. 2024. [Cross-sign language transfer learning using domain adaptation with multi-scale temporal alignment](#). *Multimedia Tools and Applications*, 83(12):37025–37051.
- Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. [Sign language transformers: Joint end-to-end sign language recognition and translation](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10033.
- Anal Roy Chowdhury and Debarshi Kumar Sanyal. 2025. [Enhancing Indian Sign Language translation via motion-aware modeling](#). In *Proceedings of the Workshop on Sign Language Processing (WSLP)*, pages 39–46.
- Pedro Dal Bianco, Oscar Stanchi, Facundo Quiroga, and Franco Ronchetti. 2025. [Multilingual sign language translation with unified datasets and pose-based transformers](#). In *Proceedings of the Workshop on Sign Language Processing (WSLP)*, pages 31–36.
- Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. [How2Sign: A large-scale multimodal dataset for continuous American Sign Language](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2735–2744.

- Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, Karen Livescu, and Alexander H Liu. 2025. [SHuBERT: Self-supervised sign language representation learning via multi-stream cluster prediction](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28792–28810.
- Youngjoon Jang, Haran Raajesh, Liliane Momeni, Gül Varol, and Andrew Zisserman. 2025. [Lost in translation, found in context: Sign language translation with contextual cues](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. 2020. [Whole-body human pose estimation in the wild](#). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 196–214.
- Abhinav Joshi, Susmit Agrawal, and Ashutosh Modi. 2023. [ISLTranslate: Dataset for translating Indian Sign Language](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10466–10475.
- Abhinav Joshi, Romit Mohanty, Mounika Kanakanti, Andesha Mangla, Sudeep Choudhary, Monali Barabate, and Ashutosh Modi. 2024. [iSign: A benchmark for Indian Sign Language processing](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10827–10844.
- Abhinav Joshi, Vaibhav Sharma, Sanjeet Singh, and Ashutosh Modi. 2025. [POSESTITCH-SLT: Linguistically inspired pose-stitching for end-to-end sign language translation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13845–13864.
- Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. 2025. [Uni-Sign: Toward unified sign language understanding at scale](#). In *International Conference on Learning Representations*.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Ubaweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. [MediaPipe: A framework for building perception pipelines](#). *arXiv preprint arXiv:1906.08172*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.
- Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 2016. [300 Faces in-the-wild challenge: database and results](#). *Image and Vision Computing*, 47:3–18.
- Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. 2022. [Open-domain sign language translation learned from online video](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6365–6379.
- Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. [Sign2GPT: Leveraging large language models for gloss-free sign language translation](#). In *International Conference on Learning Representations*.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.
- Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. [Spatial temporal graph convolutional networks for skeleton-based action recognition](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Ulrike Zeshan. 2003. [Indo-Pakistani Sign Language grammar: A typological outline](#). *Sign Language Studies*, 3(2):157–212.
- Ulrike Zeshan. 2004. [Interrogative constructions in signed languages: Crosslinguistic perspectives](#). *Language*, 80(1):7–39.
- Biao Zhang, Garrett Tanzer, and Orhan Firat. 2024. [Scaling sign language translation](#). In *Advances in Neural Information Processing Systems*, volume 37.
- Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. [Gloss-free sign language translation: Improving from visual-language pretraining](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20871–20881.
- 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*, pages 1316–1325.

A Training and Checkpoint Details

Source checkpoints. We pretrain nothing. All three sources are the public pose-only translation releases accompanying Li et al. (2025): `how2sign_pose_only_slt.pth`, `openasl_pose_only_slt.pth`, and `cs1_daily_pose_only_slt.pth`. They share a stage-1 pose pretraining on CSL-News and differ in the stage-2 translation corpus, so a comparison between two of them isolates stage 2. Each stores 627 tensors: 343 belong to the sign encoder (6.9M parameters) and 284 to the mT5-base decoder, which holds the remainder. Under 1% of the parameters touch the sign input.

Fine-tuning. Every condition trains for 20 epochs on the ISL training split with AdamW (Loshchilov and Hutter, 2019), $\beta = (0.9, 0.98)$, weight decay 0.01 applied to non-bias non-norm parameters only, one-cycle schedule with peak learning rate 3×10^{-4} and 200 warmup steps, label smoothing 0.2, at most 256 frames per clip, and bf16 autocast. Effective batch is 8 throughout; the micro-batch is 2 with gradient accumulation 4 on the 16 GB card and 4 with accumulation 2 on the 24 GB card, which changes sequence padding and is why arms are compared only within a configuration. We take the final epoch: no condition uses development-set checkpoint selection, so none benefits from a model-selection step the others did not receive. Seeds are 13, 21, and 37.

What each condition loads. Every condition instantiates the same architecture and differs only in which tensors are initialised from a source checkpoint and which are trainable (Table 9).

Per-condition chrF. Encoder + decoder 21.80; encoder only 15.02; decoder only 15.50; hand streams only 15.14; body only 15.21; face only 15.39; scratch 15.29. These are means over the same runs as Table 1.

Evaluation. Beam search, 4 beams, 100 new tokens maximum, matching the reference implementation. Scores are computed with `sacrebleu` (Post, 2018) on the full test split. Decoding under bf16 autocast is not bit-deterministic, so a fixed checkpoint varies by up to ± 0.03 BLEU-4 between passes; tables built from one decode pass carry their own control.

Larger-data condition. The data-scale comparison in Section 4.8 uses a second corpus draw

Condition	Loaded	Trained
Encoder + decoder	all 627	all
Encoder only (frozen)	343 encoder	decoder + projection
Decoder only	284 mT5	all
Hand / body / face	that stream’s 85 (170 for both hands)	all
Neither (scratch)	none	all
CSL enc + English dec	343 from CSL, 284 from How2Sign	all

Table 9: Parameter initialisation per condition. Tensors not loaded keep the architecture’s own pretrained initialisation, which for the decoder is mT5-base (Xue et al., 2021). “Frozen” means the four ST-GCN streams receive no gradient; the projection into the decoder stays trainable, since a frozen projection would confound representation quality with an unadapted interface.

of 28,512 training clips and a subsample of it at 19,099, both evaluated on the same 3,534-clip test split with the same split assignment, so only training size varies. The subsample is drawn by source video rather than by clip, so it does not split a video across the two arms. This is a different draw from the 19,093-clip split used everywhere else in the paper, and the two numbers should not be confused.

Configuration across Table 1. The encoder+decoder, frozen-encoder and scratch rows were trained at micro-batch 2 with accumulation 4; the decoder-only and single-stream rows at micro-batch 4 with accumulation 2. Effective batch is 8 in both. Where we can measure the difference, on the ASL arm of Table 2, it is worth +0.14 BLEU-4 — in the direction that would *raise* the single-stream rows, and an order of magnitude below the gap those rows are read against.

Provenance. Every run above is recorded with a manifest giving its source checkpoint, loaded tensor count, dataset, seed, and full argument list; these are reproduced from the manifests rather than from the launch scripts. Two exceptions: the manifests for one scratch seed and one frozen-encoder seed were lost with the machine they ran on, and their configuration is taken from the launcher that produced them, whose defaults match the verified manifests of their sibling runs. Both contribute only a second or third seed to a row, never a reported comparison on their own.

B Split Construction and Leakage Audit

iSign clip identifiers follow three conventions, and a rule that handles only the first two silently fragments videos of the third kind into several:

Shape	Video id ends at	Clips
<id>--<n>	final --	21,243
<hex>--<n>	final -	2,181
<id>_<suffix>	final _	497

The third shape carries hyphens inside the identifier itself, for example `Miy-F4TYnso_d`. Our original rule mis-parsed these, which is how 22 source videos came to cross the train-test boundary. Under the corrected rule the three splits hold 3,198, 411, and 413 source videos. Residual leakage is 22 videos and 22 test clips (0.90%); for the data-scale pair it is 39 of 3,534 (1.10%) and 48 of 3,534 (1.36%).

Removing the affected clips changes BLEU-4 by +0.02, +0.03, and +0.02 on seeds 13, 21, and 37 respectively. Scored alone, the leaked clips reach 1.23, 0.69, and 0.73 BLEU-4, well below the full-test-set scores of 3.58, 3.77, and 3.47. The leakage does not inflate any reported number.

C Provenance of Model Output

If the output is not retrieved from training, where does it come from? Comparing output n -grams against the training corpus, and controlling for the fact that real references share that corpus and overlap it by default, hypothesis trigrams appear in training 46.5% of the time against 20.8% for genuine references; at four-grams, 20.2% against 5.6%. The model composes from training material at a higher rate than a real translation does, yet 80% of its four-grams appear nowhere in training: recombination under weak conditioning, not verbatim retrieval.

D Test-Set Anchors

System	BLEU-4	chrF
Constant (commonest train sentence)	0.00	0.21
Random training sentence	0.06	12.63
Length-matched random sentence	0.15	15.14
Scratch (trained, no transfer)	0.46	15.29
ASL transfer (ours)	3.61	21.80

ASL transfer is a three-seed mean and scratch a two-seed mean, matching Table 1; both are rounded to two decimals here.

E Full Articulator Masking Lattice

All sixteen subsets, seed-13 checkpoint, one decode pass. Date recall is over the 111 test clips whose reference contains a date.

Streams retained	BLEU-4	chrF	Date
body + left + right + face	3.61	21.91	68.5%
body + right + face	2.90	20.56	62.2%
body + left + right	2.08	19.45	60.4%
body + right	1.26	17.42	55.0%
body + left + face	1.09	15.61	0.0%
left + right + face	0.47	15.60	27.9%
body + face	0.24	13.54	0.0%
body + left	0.22	13.83	0.0%
right + face	0.21	14.99	15.3%
left + right	0.12	14.25	0.9%
right	0.10	13.97	0.0%
body	0.05	13.42	0.0%
left + face	0.05	14.50	0.0%
left	0.03	14.05	0.0%
face	0.03	13.73	0.0%
none (blind)	0.03	13.96	0.0%

At sizes two and three, every subset containing body outperforms every same-sized subset without it. At size one the ordering reverses: body alone (0.05) is indistinguishable from blind decoding while the right hand alone reaches 0.10, which is the not-sufficient half of the claim in Section 4.6. Note also that the four subsets recovering more than half of dates all contain body together with the right hand.