

What Actually Transfers? Diagnosing Output-Projection Initialization in Low-Resource Sign Language Translation

Kai-Yang Peng Chunki Lim Shu-Kai Hsieh

Graduate Institute of Linguistics, National Taiwan University
{r14142015, r14142001, shukaihsieh}@ntu.edu.tw

Abstract

Released sign-to-text checkpoints combine visual and text-side parameters, so intact-checkpoint comparisons cannot identify what drives transfer. We study this confound in low-resource Taiwan Sign Language (TSL) by separating and recombining the visual and mT5 components of Uni-Sign checkpoints trained on Chinese Sign Language (CSL)-to-Chinese and American Sign Language (ASL)-to-English data, then fine-tuning the combinations identically on TSL. The CSL-derived visual initialization improves content-word accuracy by 15.0 points with its released mT5 parameters, but changes accuracy by only -0.5 points with untouched mT5-base, showing that the measured benefit of visual pretraining depends on the text-side initialization. We assess video dependence by replacing target clips with duration-matched held-out clips at evaluation. Further interventions localize much of the accuracy and video-dependence effects to the untied output projection: replacing an ASL-to-English model’s projection with mT5-base’s improves accuracy by 25.0 points and increases sensitivity to clip replacement. Permutation and timing controls show that the effect depends on pre-trained row-to-token correspondence and on the projection being present during fine-tuning. Absolute translation quality remains low, so we interpret these experiments diagnostically. The results show that released checkpoints should be analyzed as structured initializations whose components can interact, rather than treated as indivisible evidence of cross-language transfer.

1 Introduction

Transfer studies in low-resource sign language translation (SLT) often rely on available pretrained checkpoints as starting points. A released checkpoint, however, combines parameters from both the signed-video and text-generation components, together with the optimization history that jointly shaped them. If the source sign language also

covaries with the written output language across checkpoints, comparisons between intact models cannot separate these sources. This problem arises in multilingual and cross-language SLT (Asasi et al., 2026; Yin et al., 2022; Tanzer and Zhang, 2025) and in systems that combine separately pretrained visual and text-generation components (Zhang et al., 2023; Zhou et al., 2023; Wong et al., 2024).

Taiwan Sign Language (TSL) provides a particularly clear setting for examining this confound. The available TSL corpus pairs signed utterances with Chinese translations, while TSL is historically linked to Japanese Sign Language rather than Chinese Sign Language (CSL) (Smith, 2005; Fischer and Gong, 2010; Su and Tai, 2009; Sasaki, 2007). Uni-Sign (Li et al., 2025) releases checkpoints with the same architecture for CSL-to-Chinese and ASL-to-English, but not for TSL. Adapting one of these checkpoints therefore changes the provenance of the visual and text-side components together. An apparent transfer advantage could reflect the visual initialization, the text-side initialization, or their interaction.

We separate these possibilities by decomposing each checkpoint into the parts shown in Figure 1. We call the pose encoder and its linear bridge to mT5 the *visual initialization*. The *text-side initialization* comprises the remaining mT5 parameters, including the *output projection* that maps final hidden states to vocabulary logits. In standard mT5 this projection is tied to the input embedding, but the released Uni-Sign checkpoints store and train it as a separate `lm_head` matrix. We label each initialization by its source checkpoint. Thus, the “CSL-Daily text-side initialization” denotes the mT5 portion of the CSL-Daily CSL-to-Chinese checkpoint (Zhou et al., 2021); “mT5-base” denotes the published multilingual model before sign-language training. In projection swaps, the *recipient* supplies the rest of the model and the *donor*

supplies only the substituted output projection.

Our intervention strategy follows work on identifying transferable components (Neyshabur et al., 2020), model stitching and parameter editing (Bansal et al., 2021; Ilharco et al., 2023), output-embedding specialization (Chung et al., 2021), and cross-lingual lexical reinitialization (Artetxe et al., 2020; de Vries and Nissim, 2021; Minixhofer et al., 2022; Dobler and de Melo, 2023). Few-sample fine-tuning motivates explicit checks for optimization sensitivity (Zhang et al., 2021), and signed-language NLP motivates cautious evaluation (Yin et al., 2021; Müller et al., 2023). Unlike WECHSEL or FOCUS, we keep the vocabulary, tokenizer, architecture, and TSL adaptation task fixed while varying the parameters used for initialization.

We proceed from coarse to fine interventions. First, we test whether a fixed visual initialization contributes equally under different text-side initializations. We then isolate the output projection and use random, rescaled, and row-permuted controls to identify which of its properties matter. Finally, we vary replacement timing and learning rate to distinguish an initialization effect from a better final readout or a fixed checkpoint ranking. Accordingly, our causal interpretation is limited to output-projection initialization under low-resource adaptation. The experiments do not isolate Chinese versus English output-language pretraining as the causal factor.

2 Setup

Data and splits. We use the Taiwan Sign Language Corpus (Ministry of Culture, Taiwan, 2025), which contains 407 filmed paragraphs, 5,282 annotated utterances with Chinese translations, and 25 signers. We split at the paragraph level because adjacent utterances often share a signer, topic, and material from the same clause. Each of three folds uses 243–245 paragraphs (3,176–3,255 utterances) for training and 135–137 (1,760–1,761) for testing. Appendix G gives the per-fold counts. The split is not signer-disjoint: at most 1 test signer per fold is absent from training. The reported results therefore measure generalization to held-out *paragraphs*, not held-out *signers*.

Forced-choice diagnostic. BLEU is near the floor for every model in this study, so we use a forced-choice diagnostic that remains informative when free translation quality is low. It contains 940 four-way items spanning 238 content

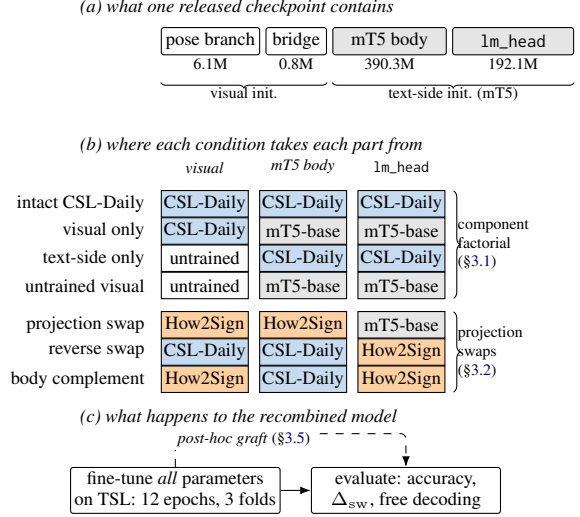


Figure 1: Intervention protocol. (a) A released UniSign checkpoint is partitioned into a pose branch, the bridge into mT5, the mT5 body, and the untied output projection (1m_head). (b) Each row specifies the source checkpoint for each component. The first four rows show the core 2×2 component factorial; §3.1 adds the shared stage-1 visual initialization as a provenance control. The final three isolate the output projection and its complement in §3.2. (c) All recombined models are fine-tuned on TSL and evaluated under the same procedure. The dashed path marks the post-hoc replacement in §3.5, which installs the projection only after fine-tuning.

words. Each item replaces one content-word slot in a reference sentence with the gold candidate or one of three distractors. Distractors match the gold in corpus part of speech, are close in log frequency, and avoid exact location-and-handshape collisions (Appendix A). For a candidate c tokenized as $y_1, \dots, y_T$ and a video v , we score each token conditional on the candidate’s own preceding tokens and define

$$s(c|v) = \frac{1}{T} \sum_{t=1}^T \log p_{\theta}(y_t | y_{<t}, v). \quad (1)$$

The model selects $\arg \max_c s(c|v)$. Conditioning on the candidate prefix avoids dependence on previous model predictions; mean log probability controls for candidate length. Absolute accuracy is not by itself evidence of visual grounding. One rank-balanced Qwen2.5-1.5B-Instruct mean-token baseline is exactly 25% (Qwen Team, 2024), while other text-only scorers given the sentence without video reach 38.0–55.3%. We therefore use accuracy as a task-level measure and base claims about video dependence on paired input interventions.

Dependence on clip identity. We measure whether predictions depend on the target video by replacing that video with a different, duration-matched clip from another held-out paragraph while leaving the sentence and candidate set unchanged. Let A_{target} and A_{swap} denote accuracy with the original and substituted clips. We define *wrong-clip sensitivity* as $\Delta_{\text{sw}} = A_{\text{target}} - A_{\text{swap}}$. Positive values therefore indicate that predictions benefit from the identity of the target clip. We prefer a real substitute to an all-zero pose sequence because the latter lies outside the training distribution and can even reverse the measured effect (Appendix A). A near-zero Δ_{sw} should not be interpreted as evidence that no visual information is represented internally. It can also arise when a decoder has collapsed to nearly constant output. We therefore interpret Δ_{sw} together with free decoding and the bidirectional parameter interventions. A second 975-item diagnostic targeting referential pointing signs provides an independent check (Appendix A).

In tables, Acc. denotes A_{target} , while A_{swap} is the mean accuracy over three duration-matched substitute clips. Accuracy contrasts in the text subtract A_{target} between model cells; factorial interactions subtract those contrasts. Contrasts use paired unrounded scores, so subtraction of displayed one-decimal cells can occasionally differ by 0.1 point from the reported contrast.

Models, training, and statistics. In Uni-Sign, the visual initialization comprises a 6.1M-parameter pose branch over COCO-WholeBody keypoints (Jin et al., 2020; Jiang et al., 2023) plus a 0.8M-parameter bridge into mT5-base (Xue et al., 2021). We use released CSL-Daily, CSL-News, How2Sign, OpenASL, and WLASL checkpoints (Zhou et al., 2021; Li et al., 2025; Duarte et al., 2021; Shi et al., 2022; Li et al., 2020). All inherit stage-1 CSL-News pose pretraining, so the “CSL-Daily visual initialization” includes that shared stage plus later CSL-Daily adaptation (§3.1). The untied $250,112 \times 768$ output projection contains 192.1M parameters and is stored separately from the input embedding, allowing independent replacement. RAND-VIS, the no-visual-pretraining baseline, pairs an untrained pose branch with untouched mT5-base.

Unless noted otherwise, conditions are fine-tuned end to end for 12 epochs with Uni-Sign’s published learning rates over three paragraph folds; key

visual init.	CSL-Daily		mT5-base	
	Acc.	Δ_{sw}	Acc.	Δ_{sw}
CSL-Daily (stage 1 + adapt.)	56.5	+22.1	39.6	-0.5
CSL-News (stage 1 only)	56.0	+22.2	40.2	+1.3
untrained	41.5	+1.6	40.1	+0.0

Table 1: Factorial comparison, pooled over three folds. Acc. is target-clip accuracy; Δ_{sw} is the percentage-point drop under duration-matched real-clip substitution. The middle row isolates shared stage-1 CSL-News pose pretraining from later CSL-Daily adaptation.

projection replacements are rerun with three seeds on fold 0. Confidence intervals use paragraph-clustered bootstrap resampling and paired contrasts use paragraph-level permutation tests. Full hyperparameters and checkpoint provenance are reported in Appendix G.

3 Results

We first establish whether visual and text-side initializations interact, then localize that dependence and test what makes the output projection useful. We finally test generality, replacement timing, free generation, and optimization sensitivity.

3.1 Visual pretraining depends on the text-side initialization

Table 1 reports cell-level Acc. and Δ_{sw} . The contrasts here use Acc., not Δ_{sw} . With the CSL-Daily text-side initialization, the visual-pretraining contrast is $56.5 - 41.5 = +15.0$ points ($[+12.0, +17.9]$); with mT5-base it is $39.6 - 40.1 = -0.5$ points ($[-2.2, +1.2]$). Their difference is the interaction: $+15.0 - (-0.5) = +15.5$ points ($[+12.1, +18.9]$, $p < 0.001$). The corresponding Δ_{sw} values are +22.1 versus +1.6 in the first column and -0.5 versus $+0.0$ in the second, showing the same conditional pattern in clip dependence.

The middle row locates the accuracy interaction earlier in Uni-Sign training. Stage-1 pose pretraining contributes $56.0 - 41.5 = +14.5$ points with the CSL-Daily text-side initialization, while later CSL-Daily adaptation adds only $56.5 - 56.0 = +0.5$ points. With mT5-base, stage-1 pretraining contributes $40.2 - 40.1 = +0.1$ points. The conditional benefit is therefore already present after shared pose pretraining rather than being specific to CSL-Daily visual adaptation.

Cross-loading the two Chinese-output checkpoints leaves both combinations near their intact scores (56.0% and 56.0%). Thus the interaction

does not require the original jointly trained pairing, motivating a finer intervention on the text-side initialization.

3.2 The interaction localizes to the output projection

We use parameter displacement only to choose the first text-side component to intervene on. Relative to mT5-base, released input embeddings differ by only 0.0015–0.0025 in relative L_2 , whereas the released output projections differ by 0.92–0.98. These distances are descriptive and are not treated as causal explanations.

Replacing only How2Sign’s output projection with the mT5-base projection raises content-word accuracy by 25.0 points ([+21.2, +28.9]) and increases wrong-clip sensitivity by +14.7 points ([+11.7, +17.7], $p < 0.001$). The reverse intervention has the opposite effect: replacing CSL-Daily’s projection with How2Sign’s lowers accuracy from 56.5% to 24.3% and reduces Δ_{sw} from +22.1 to +0.3 points. Its residual wrong-clip effect differs from RAND-VIS by only +0.3 points ([−1.0, +1.6], $p = 0.668$); the interval includes zero, so these data do not show a detectable separation between the two conditions. Both directions replicate across three seeds.

A complement control asks whether the surrounding mT5 parameters can reproduce the same effect without changing the projection. How2Sign receives all CSL-Daily mT5 parameters except the output projection while retaining its own projection. This complement contains 390.3M distinct parameters, about twice the 192.1M parameters in the projection, but reaches only 25.5% accuracy with $\Delta_{sw} = +2.1$ and remains degenerate under free decoding. The much larger surrounding component therefore does not reproduce the rescue. We consequently localize much of the text-side dependence to the output projection, without treating it as the only parameter that can matter.

3.3 Token-aligned rows are necessary for a useful projection

Localization to the output projection still leaves several explanations open. The successful donors differ from the original How2Sign projection in whether they are pretrained, their overall scale, their script-specific row statistics, and the correspondence between vocabulary rows and the tokens scored by those rows. Resetting the projection gives a modest +3.4-point gain over How2Sign’s

Recipient	output-projection donor	Acc.	Δ_{sw}
How2Sign	own	23.9	+0.5
	random, norm-matched	27.3	+2.0
	mT5-base, rows permuted	27.4	+2.0
	CSL-Daily	48.5	+14.3
	mT5-base	48.9	+15.2
OpenASL	own	24.6	+0.7
	CSL-Daily	49.7	+17.3
	mT5-base	55.7	+22.6
WLASL	own	26.4	+1.8
	CSL-Daily	45.9	+10.8
	mT5-base	50.0	+15.6
CSL-Daily	own	56.5	+22.1
	How2Sign	24.3	+0.3
	mT5-base	63.0	+29.0

Table 2: Output-projection replacements at initialization. The recipient supplies all parameters except the output projection, and the named donor supplies only that matrix. All other parameters and the training procedure are held fixed. Columns report content-word accuracy and wrong-clip sensitivity in percentage points; every row pools the same three folds. Contrasts quoted in the text use paired, unrounded scores.

original projection ([+0.4, +6.3]), so removing that initialization can help. The norm-matched random projection nevertheless remains 21.6 points below mT5-base ([−25.7, −17.5]). Changing the random initialization’s scale changes accuracy by only +0.4 points ([−1.6, +2.4]), while rescaling How2Sign’s own projection changes accuracy by at most about one point (Appendix B).

The row-permutation control directly tests token correspondence. Before fine-tuning, we randomly permute the vocabulary rows of the mT5-base projection. This preserves the exact multiset of rows, its Frobenius norm, row-norm distribution, singular values, and conditioning, while changing which token each pretrained row scores. The input embedding is left unchanged. If the recipient needed only generic pretrained matrix structure, the permuted donor should perform like the intact donor.

Instead, the permuted projection reaches 27.4%, compared with 48.9% for the same matrix without permutation. The 21.5-point loss is large ([−25.5, −17.5]), and wrong-clip sensitivity falls to +2.0 points. Relative to the norm-matched random projection, the estimated difference is only +0.1 points ([−2.7, +2.9]); relative to How2Sign’s original projection it is modestly positive ([+0.8, +6.3]). Permuting rows only within script groups also fails to recover the intact donor (24.8%, [−29.7, −16.7]). The useful property is therefore not global pre-

trained structure or script-level scale. The rows must remain aligned with the tokens they were pre-trained to score.

This does not identify lexical semantics as the mechanism, because row-to-token correspondence remains confounded with other token-specific mT5 geometry. Consistent with that narrower interpretation, a gloss-to-Chinese projection trained without video also improves How2Sign, and improvement does not track video provenance or distance from mT5-base.

3.4 The effect replicates across recipients

The same three-fold replacement pattern appears on OpenASL and WLASL (Table 2). The mT5-base projection improves content-word accuracy by 31.2 and 23.6 points, respectively ([+27.4, +35.0]; [+19.3, +28.0]); CSL-Daily’s projection yields gains of 25.1 and 19.5 points, with wrong-clip sensitivity rising from near zero in both. The effect is therefore not How2Sign-specific, although our evidence remains within the Uni-Sign architecture.

3.5 The projection must be present during adaptation

A timing intervention tests whether the donor is merely a better final readout. Installing CSL-Daily’s projection before fine-tuning improves How2Sign by 24.6 points, whereas replacing the projection in the fully trained model changes accuracy by only -1.5 points. Across three folds, the post-hoc graft leaves wrong-clip sensitivity near zero (+0.3 points, $[-0.9, +1.4]$, $p = 0.673$).

A 768×768 ridge map fitted on held-out development states provides a stronger readout control. It lowers donor cross-entropy below How2Sign’s own projection and recovers 11.8 accuracy points, yet Δ_{sw} remains near zero. Thus a better final linear readout can improve lexical scoring without reproducing the training-time change in clip dependence. The main effect requires the donor to be present during adaptation (Appendix C).

3.6 Generation and optimization provide complementary evidence

Unconstrained decoding shows the same ordering. How2Sign with its original projection obtains BLEU 0.01 and only 2.4% distinct outputs, indicating near-constant generation. With mT5-base it reaches BLEU 2.07 and 77.3% distinct outputs. A norm-matched random projection reaches BLEU 0.11 with 11.6% distinct outputs, an intermediate

stock-fragment regime (Appendix F). CSL-Daily improves from BLEU 1.93 to 4.38 with mT5-base, whereas the How2Sign projection yields BLEU 0.00 with a 93.9% modal output. The replacement can therefore prevent generation collapse without producing high-quality TSL translation.

Equation 1 can remain informative when autoregressive decoding collapses because it scores each candidate given its own prefix. A learning-rate sweep provides a separate optimization check: at 3×10^{-4} , intact How2Sign differs from RAND-VIS by -1.2 points ($[-7.6, +4.8]$), while the projection effect remains 25.2 points ($[+19.8, +30.7]$). Appendix D reports the full sweep.

4 Discussion and conclusion

A fixed visual initialization is strongly beneficial with the CSL-Daily text-side initialization and nearly neutral with mT5-base, showing why intact checkpoints can obscure component interactions. The stage-1 condition further shows that this conditional benefit is already present after the shared CSL-News pose-pretraining stage, before CSL-Daily-specific visual adaptation. Because that stage itself uses CSL-News, this locates the interaction without establishing language-general visual pretraining. Much of the dependence localizes to the output projection: one-matrix replacements improve three ASL-derived recipients, the reverse replacement damages CSL-Daily, and the twice-larger complement does not reproduce the effect. Row permutation and timing show that token-aligned pretrained rows must be present during adaptation; learning-rate sensitivity shows that initialization is not the whole account.

More broadly, when checkpoint components have different training histories, the checkpoint name is too coarse a unit of analysis. Transferability is better treated as a property of a component in its initialization context than in isolation. This separates what a component carries into adaptation from whether the surrounding initialization lets it become useful. Future checkpoint comparisons should therefore report component provenance and, where feasible, include recombination controls before attributing differences to source-language transfer. These methodological claims do not establish language matching, semantic grounding, or whether generation collapse causes low wrong-clip sensitivity.

Limitations

The evidence comes from one TSL corpus, one Uni-Sign/mT5 architecture, and one low-resource adaptation setting. Checkpoint provenance and written output language are not independently varied across the released checkpoints, so the experiments do not identify a Chinese-versus-English language effect. The untied 192.1M-parameter output projection is also specific to this architecture.

Our splits are constructed at the paragraph level rather than by signer. At most 1 test signer per fold is absent from training, so the evaluation measures generalization to held-out paragraphs and not to unseen signers. Whether the same effects hold under a genuine cross-signer split remains untested and would be necessary for stronger claims about deployment.

The row-permutation experiment establishes the importance of row-to-token correspondence but does not identify which token-specific property is causal. Lexical-semantic organization and other geometry learned during mT5 pretraining remain entangled within that correspondence. The experiment therefore narrows the mechanism without fully specifying it.

The forced-choice diagnostic also retains substantial text-side signal. Matched text-only scorers reach 38.0–55.3%, and a scorer with additional discourse context reaches 66.2%. Moreover, generation collapse can itself reduce measured wrong-clip sensitivity, so the diagnostic does not fully separate these phenomena. BLEU remains low even after generation recovers. Some auxiliary donor controls—including text-only and within-script permutation variants—as well as learning-rate cells and mechanistic probes use fold 0 only; the main random and full-permutation controls now pool all three folds. A fold-0 label-smoothing check does not rescue intact How2Sign (Appendix D), but other hyperparameters are not systematically ablated.

Ethical Considerations

The Taiwan Sign Language Corpus ([Ministry of Culture, Taiwan, 2025](#)) is a research-use beta release. We redistribute neither corpus material nor signer metadata, and all reported split statistics are aggregate counts. Pose keypoints can retain identity-related cues, so we handle them as potentially identifying data. The location-and-handshape filter used in the diagnostic is an automated corpus

check rather than linguistic validation. It cannot replace review by fluent TSL signers or participation by Deaf communities ([Yin et al., 2021](#); [Müller et al., 2023](#)). The diagnostic design and analysis did not include consultation with Deaf signers, and we therefore make no claims about usability or community benefit.

References

- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. [On the cross-lingual transferability of monolingual representations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.
- Sobhan Asasi, Ozge Mercanoglu Sincan, and Richard Bowden. 2026. SIGNET: Motion-level knowledge transfer for cross-language sign language translation. In *European Conference on Computer Vision (ECCV)*. To appear. arXiv:2606.28626.
- Yamini Bansal, Preetum Nakkiran, and Boaz Barak. 2021. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pages 225–236.
- Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. [Rethinking embedding coupling in pre-trained language models](#). In *International Conference on Learning Representations (ICLR)*.
- Wietse de Vries and Malvina Nissim. 2021. [As good as new. How to successfully recycle English GPT-2 to make models for other languages](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 836–846, Online. Association for Computational Linguistics.
- Konstantin Dobler and Gerard de Melo. 2023. [FOCUS: Effective embedding initialization for monolingual specialization of multilingual models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13440–13454, Singapore. Association for Computational Linguistics.
- Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giró-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 (CVPR)*, pages 2735–2744.
- Susan Fischer and Qunhu Gong. 2010. Variation in East Asian sign language structures. In Diane Brentari, editor, *Sign Languages*, pages 499–518. Cambridge University Press, Cambridge.

- GLM Team. 2024. [ChatGLM: A family of large language models from GLM-130B to GLM-4](#). *Preprint*, arXiv:2406.12793.
- Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hananeh Hajishirzi, and Ali Farhadi. 2023. [Editing models with task arithmetic](#). In *International Conference on Learning Representations (ICLR)*.
- Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. 2023. RTM-Pose: Real-time multi-person pose estimation based on MMPose. *arXiv preprint arXiv:2303.07399*.
- 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.
- 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.
- 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 (ICLR)*.
- Ministry of Culture, Taiwan. 2025. Taiwan Sign Language Corpus. Beta release, research use. Accessed 2026-08-05. <https://tslcorpus.moc.gov.tw/>.
- Benjamin Minixhofer, Fabian Paischer, and Navid Rekasbas. 2022. [WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.
- Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. [Considerations for meaningful sign language machine translation based on glosses](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.
- Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. 2020. What is being transferred in transfer learning? In *Advances in Neural Information Processing Systems*, volume 33, pages 512–523.
- Qwen Team. 2024. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Daisuke Sasaki. 2007. Comparing the lexicons of Japanese Sign Language and Taiwan Sign Language: A preliminary study focusing on the difference in the handshape parameter. In David Quinto-Pozos, editor, *Sign Languages in Contact*, pages 123–150. Gallaudet University Press, Washington, DC.
- 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Wayne H. Smith. 2005. Taiwan Sign Language research: An historical overview. *Language and Linguistics*, 6(2):187–215.
- Shiou-fen Su and James H.-Y. Tai. 2009. Lexical comparison of signs from Taiwan, Chinese, Japanese, and American Sign Languages: Taking iconicity into account. In James H.-Y. Tai and Jane Tsay, editors, *Taiwan Sign Language and Beyond*, pages 149–176. The Taiwan Institute for the Humanities, National Chung Cheng University, Chia-Yi, Taiwan.
- Garrett Tanzer and Biao Zhang. 2025. [YouTube-SL-25: A large-scale, open-domain multilingual sign language parallel corpus](#). In *International Conference on Learning Representations (ICLR)*.
- 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 (ICLR)*.
- 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, Online. Association for Computational Linguistics.
- Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. 2022. MLSLT: Towards multilingual sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5109–5119.
- Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. [Including signed languages in natural language processing](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.
- Biao Zhang, Mathias Müller, and Rico Sennrich. 2023. SLTUNET: A simple unified model for sign language

translation. In *International Conference on Learning Representations (ICLR)*.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. 2021. [Revisiting few-sample BERT fine-tuning](#). In *International Conference on Learning Representations (ICLR)*.

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 (ICCV)*, 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 (CVPR)*, pages 1316–1325.

A Diagnostics and statistical procedure

The content-word diagnostic replaces one reference slot with the gold word or one of three distractors. Gold and distractor tokens share the corpus POS tag, and distractor types are selected near the gold in log frequency. Candidates must be standalone tokens under jieba and must not share the gold sign’s location and handshape signature. That filter removes exact sign-phonological collisions, but near-minimal pairs remain: an audit finds 16.6% of distractors share the gold location, 6.3% share its handshape, and 47.0% of items contain at least one such distractor. These are automated corpus checks, not human judgments.

The gold occupies each Qwen rank equally often, which makes the construction-specific Qwen mean-log-probability baseline exactly 25%. That guarantee does not generalize to other scorers. Under the same sentence-level input, Qwen with summed log probability reaches 55.3%, a GLM-family scorer (GLM Team, 2024) reaches 38.0%, and a unigram prior reaches 38.3%. With preceding sentences from the same paragraph Qwen reaches 66.2%, but the video conditions receive no such discourse context, so this is an unmatched reference.

For the wrong-clip intervention, the substituted clip comes from another held-out paragraph and is matched in duration to the target. The median duration difference is 9 ms, and results average over three substitute clips. A within-paragraph substitution, which holds signer and session fixed, still reduces CSL-Daily accuracy by 18.9 points ([+15.9, +21.8]). The effect therefore persists

when signer, session, and paragraph context are held more closely matched. We do not use an all-zero clip as the primary intervention because it lies outside the training distribution and can even reverse the direction of the measured effect.

The pointing-sign diagnostic. The second diagnostic contains 975 four-way items targeting referential pointing signs. It is constructed by the same procedure, with the varied slot placed at a referential expression rather than a content word. The items are more predictable from text alone, so absolute accuracy is less informative than on the content-word set. Their wrong-clip effects are substantially larger, however, which makes the set a useful independent check.

Every main-text intervention produces the same direction of effect on this set. Intact How2Sign scores 45.7% with $\Delta_{\text{sw}} = -1.2$. Installing mT5-base’s projection raises accuracy to 85.8% and Δ_{sw} to +40.0, compared with +15.2 on the content-word set. In the reverse direction, intact CSL-Daily scores 77.6% with $\Delta_{\text{sw}} = +33.7$. Under How2Sign’s projection it falls to 46.7% and $\Delta_{\text{sw}} = +0.0$, eliminating measured clip dependence. Frame shuffling, which removes temporal order while preserving the frame distribution, reduces intact CSL-Daily accuracy by 24.8 points and the reverse-replacement condition by 0.3.

The two item sets therefore agree on the direction and the localization of every intervention we report, differing mainly in how much of the available wrong-clip effect each resolves. §D notes the one place they disagree, which concerns additivity with the learning rate rather than the projection effect itself.

For each cell, Acc. is the proportion of diagnostic items whose gold candidate ranks first. Swap accuracy is computed for each of three duration-matched substitute clips and averaged; Δ_{sw} subtracts this mean from target-clip accuracy. Accuracy contrasts subtract paired cell accuracies on shared items, and interaction contrasts subtract two such paired differences. All main content-word confidence intervals use 10,000 bootstrap samples resampled over paragraphs; paired p -values use paragraph-level sign-flip permutation tests. Thus the paragraph, not the item, is the resampling unit. A pooled three-fold content-word contrast contains 940 items from 341 paragraphs, and the paragraph count determines the effective sample size. Holm correction is applied to the original family

of reported contrasts. The component interaction remains significant after correction, whereas the small encoder-only content-word contrast does not.

B Full donor and reset controls

How2Sign 1m_head donor	Acc.	Δ_{sw}
own	23.9	+0.5
mT5-base	48.9	+15.2
mT5-base, rows permuted	27.4	+2.0
mT5-base, permuted in script	24.8	+2.0
CSL-Daily	48.5	+14.3
CSL-News	46.4	+14.3
TSL text	49.1	+14.5
gloss→Chinese	47.3	+10.7
TSL text, drift-matched	43.0	+11.4
random, norm-matched	27.3	+2.0

Table 3: Every output-projection donor tested on the How2Sign recipient, with content-word accuracy and wrong-clip sensitivity. The CSL-News, TSL-text, gloss-to-Chinese, drift-matched, and within-script-permutation rows are fold 0 only; the norm-matched random and full-permutation rows pool three folds. Paired contrasts quoted in the text compare cells on shared items.

All neutral or Chinese-capable pretrained donors improve How2Sign, and the estimated gains overlap broadly despite their different provenance. The gloss-to-Chinese projection improves accuracy by 22.7 points ([+16.6, +29.2]). Its estimated difference from mT5-base is -0.6 points, with an interval spanning zero ($[-3.2, +2.0]$). The TSL-text donor labeled “drift-matched” was trained until its projection reached a relative L_2 distance of 0.913 from mT5-base, comparable to the released fine-tunes. It still improves How2Sign, but its development cross-entropy is 2.63. Because this displacement was obtained partly by overtraining, the donor is not a clean matched instrument.

The random reset distinguishes pretraining from mere removal of the English projection: a norm-matched random projection is 21.6 points below mT5-base ($[-25.7, -17.5]$). Changing the random initialization’s row-norm scale changes accuracy by only $+0.4$ points ($[-1.6, +2.4]$), so scale alone does not explain the gap to mT5-base. Rescaling How2Sign’s own projection before training, using CJK-row correction, radius correction, or both, changes accuracy by at most about one point and leaves wrong-clip sensitivity near zero. The same CJK correction can move accuracy when applied post-hoc, suggesting a test-time calibration effect rather than a successful training initialization.

Appendix E reports a complementary gradient probe before any parameter update.

The permutation donors are described in §3.3. Both are built from mT5-base’s own projection with the vocabulary rows reordered, so their row multiset, Frobenius norm, row-norm distribution and singular values are identical to mT5-base’s by construction. The full permutation also changes the script region associated with each row, whereas the within-script permutation does not.

C Post-hoc readout tests

A post-hoc replacement re-scores a fully trained recipient without further training. Pooled over the three folds, replacing How2Sign’s projection with CSL-Daily’s changes accuracy by -1.5 points. Across three fold-0 seeds, the corresponding changes are -0.6 , $+1.5$, and -3.6 points, with score-margin changes of -0.223 , -0.209 , and -0.207 nats. The self-replacement control changes little, and a post-hoc mT5-base projection also gives small mixed effects across three seeds. The same mT5-base projection at initialization improves How2Sign by 25.0 points and CSL-Daily by 6.5 points.

We additionally fit 768×768 interface maps on held-out development states, to test whether a donor projection simply expects a different hidden-state basis. How2Sign and CSL-Daily decoder states are indeed far apart. A ridge map lowers the CSL-Daily donor’s dev cross-entropy on How2Sign states below the recipient’s own projection and recovers about 12 points of contrastive accuracy, yet Δ_{sw} stays near zero; related orthogonal and cross-entropy-trained maps show the same dissociation. On CSL-Daily the same kind of aligned replacement leaves a reduced but nonzero wrong-clip effect. The alignment can therefore repair a lexical readout while leaving measured wrong-clip sensitivity near zero in the trained How2Sign endpoint.

D Optimization and model-selection controls

The mT5 learning-rate sweep changes the interpretation of the intact ASL baseline but not the projection effect. On fold 0, How2Sign with its own projection scores 25.2% at 10^{-4} and 39.4% at 3×10^{-4} ; with mT5-base’s projection the corresponding cells reach 56.7% and 64.5%. The paired projection effects are large at all three rates. At

3×10^{-4} , the projection still improves accuracy by 25.2 points.

Both diagnostics show that a higher learning rate can restore wrong-clip sensitivity, but they differ on whether the learning-rate and projection effects are additive. On content words the projection and learning-rate effects remain separable. On pointing signs the additional neutral-projection gain at the higher rate is +5.1 points ($[-0.4, +10.7]$, $p = 0.074$). We therefore avoid claiming a universal additive mechanism.

A fold-0 run with label smoothing set to zero leaves intact How2Sign at 23.9%, a -0.6 -point change from the default setting ($[-3.3, +2.0]$), with $\Delta_{\text{sw}} = -0.3$. Removing label smoothing alone therefore does not rescue this baseline.

The projection effect also persists when model selection is held to the final epoch. Three cells were rerun while saving epoch 12 in addition to the best-dev checkpoint. The effect is 22.7 points at the best-dev epochs and 24.8 points at epoch 12 ($[+18.0, +32.1]$). Thus the gap is not removed by evaluating both runs at the same final epoch.

E Gradient probe

Before any parameter update, we backpropagate one training step through 12 shared batches for eight recipient/projection combinations, measuring the visual-component gradient share, $\|(g_{\text{pose}}, g_{\text{bridge}})\|/\|g_{\text{total}}\|$, where g_{bridge} is the pose-to-mT5 bridge gradient. The four configurations that later collapse lie between 0.390 and 0.510; the four that later avoid collapse lie between 0.555 and 0.634. The ordering has no overlap, but it should not be read as a continuous predictor: CSL-Daily with mT5-base’s projection has a lower visual-component share than intact CSL-Daily while reaching higher final accuracy. The probe shows only that the training signal begins differently.

Absolute gradients clarify two distinct failures. How2Sign with a random projection has a very large pose gradient (34.14) but an even larger mT5-body gradient (82.62), so this share is low because the decoder requires repair. CSL-Daily with How2Sign’s projection instead reduces the pose gradient to 2.52 from 6.78 under its own projection, a small absolute gradient on the visual components. Across-batch gradient cosine stays near zero in all conditions, so this coherence measure provides no separation between the two outcome groups in

these runs.

F Free decoding and a noncollapsed control

Cell	BLEU	distinct	modal
How2Sign, own	0.01	2.4	34.2
How2Sign + mT5-base	2.07	77.3	1.7
How2Sign + permuted	0.19	25.2	8.7
How2Sign + random	0.11	11.6	15.0
How2Sign + CSL-Daily	1.36	86.4	2.0
CSL-Daily, own	1.93	85.6	3.0
CSL-Daily + mT5-base	4.38	88.6	1.3
CSL-Daily + How2Sign	0.00	1.2	93.9
How2Sign + CSL body, own proj.	0.00	0.9	95.4

Table 4: Unconstrained generation, fold 0. “Distinct” and “modal” are the percentage of output strings that are unique and the share taken by the single most frequent string. A near-zero BLEU with high distinctness and a near-zero BLEU with one repeated string are qualitatively different failure modes, which BLEU alone does not separate.

Projection	Illustrative free-decoding output
own	“我” (degenerate)
random	“我會去” (stock fragment)
mT5-base	“怎麼用手語溝通?” (multiword sentence)

Table 5: The three output regimes made concrete: hand-selected How2Sign fold-0 outputs illustrating qualitative differences that a near-zero BLEU hides. All quantitative claims rest on the full-set statistics above, never on these examples.

Table 5 illustrates the three generation regimes: near-constant output, repeated stock fragments, and multiword sentences after projection replacement.

The complement control takes CSL-Daily’s whole mT5 half except `lm_head` and places it on How2Sign. It scores 25.5% with $\Delta_{\text{sw}} = +2.1$ on the content-word diagnostic and remains degenerate under free decoding (BLEU 0.00, distinct 0.9%). The large surrounding mT5 component therefore does not substitute for the projection on either diagnostic or on generation.

To test whether the component ordering is peculiar to collapsed video translation, we replace pose features with gold gloss sequences and train the same mT5 to Chinese for 12 epochs, varying only the initial `lm_head`. The neutral projection reaches dev CE 2.050, BLEU 26.46 and 95.8% distinct outputs. The CSL-Daily projection reaches BLEU 11.99, the How2Sign and OpenASL projections

reach 1.89 and 1.79, and the norm-matched random projection reaches 1.02 with 33.1% distinct outputs. After training, the task-specific projections move only about 0.12–0.15 in relative L_2 from their own initializations. The same broad ordering therefore persists in a regime where the top model is not degenerate, and the output projection remains strongly anchored to where it started.

G Checkpoint provenance and reproducibility

Splits. The corpus holds 407 paragraphs (207 monologue, 200 dialogue), 5,282 utterances and 25 signers. Across the three folds, training takes 243–245 paragraphs (3,176–3,255 utterances), development 27 (267–345), and test 135–137 (1,760–1,761). Folds are balanced on paragraph type and diagnostic-item count by greedy largest-first assignment, with ties broken on a hash of the paragraph identifier, so they are deterministic. Signer identity is not used as a splitting variable: at most 1 test signer per fold does not also appear in that fold’s training split, so these are held-out-paragraph splits and not cross-signer ones. We report signer allocation only in aggregate, consistent with redistributing no signer metadata.

Provenance and hyperparameters. Uni-Sign initializes each downstream sign checkpoint from stage-1 pose-only pretraining on CSL-News. The released mT5 bodies stay close to this common ancestor while `lm_head` moves much further. Input embeddings shift only about 0.0015–0.0025 in relative L_2 from mT5-base, whereas released output projections shift by about 0.92–0.98. Script-wise row norms also differ between CSL and ASL checkpoints, but the random, rescaling and permutation controls in Appendix B are inconsistent with script scale as a sufficient explanation of the training-time improvement.

Unless otherwise noted, training uses AdamW, batch size 8 with two-step gradient accumulation, a pose-branch learning rate of 3×10^{-4} , an mT5 learning rate of 3×10^{-5} , 5% warmup and label smoothing 0.2. Clips are capped at 192 frames and targets at 64 tokens. The learning-rate sweep changes only the mT5 rate. Cells are selected by development cross-entropy, with the final-epoch control of Appendix D. All reported distractor candidates are unseen as diagnostic labels during training. Code, analysis scripts and the per-run measurements behind every number re-

ported here are available at <https://github.com/lopentu/wslp-transfer-artifact>. The diagnostic items are released as metadata and construction scripts rather than as text: no corpus material or signer metadata is redistributed, and no checkpoint weights are, since the donors carry no redistribution license.