Zero-Shot Cross-Lingual Recognition of Sign Language Handshapes

Universitat Pompeu Fabra Barcelona

Marcel Granero-Moya* Carolina del Corral Farrarós*
Gloria Haro Coloma Ballester Ricardo Marques

Universitat Pompeu Fabra, Barcelona * equal contribution marcel.granero@upf.edu

Project page Paper

Motivation

- **Over 200 sign languages**, yet **phonological annotations exist for almost none**, Catalan Sign Language (LSC) among them.
- Handshapes can be decomposed into **five phonological features shared across languages** [1]: selected fingers, finger flexion, spread, thumb position, thumb contact.
- Goal: **zero-shot LSC handshape recognition through phonological features** predicted by models trained on American Sign Language (ASL) only, **no LSC video or label in training**.

Data and models

Dataset	Role	Handshapes	Glosses	Signers	Videos	Recording format	Aspect ratio	Frame rate
Sem-Lex [2]	ASL train & test	58	3.1k	41	85k	webcam	4:3	∼30 fps
PopSign [3]	ASL train & test	37	250	47	175k	smartphone	3:4	∼30 fps
LSC-Benchmark	LSC zero-shot test	37	–	1	74	studio	16:9	50 fps

Table 1. Datasets. ASL phonological labels are mapped through the glosses with ASL-LEX 2.0 [4]. The LSC-Benchmark is a pre-release subset with only two clips per handshape.

Three model architectures trained to predict phonological features, separately on Sem-Lex and on PopSign:

- **MLP** [5]: three fully connected layers on the 21 MediaPipe 3D hand landmarks of one least-motion frame of the dominant hand.
- **SL-GCN** [6]: spatio-temporal graph over MediaPipe pose sequences extracted from the raw videos; isolated-sign pre-training, then joint gloss and phonological features fine-tuning.
- **SHuBERT** [7]: multimodal transformer pre-trained on YouTube-ASL, fully fine-tuned for phonological feature prediction.

Zero-shot pipeline

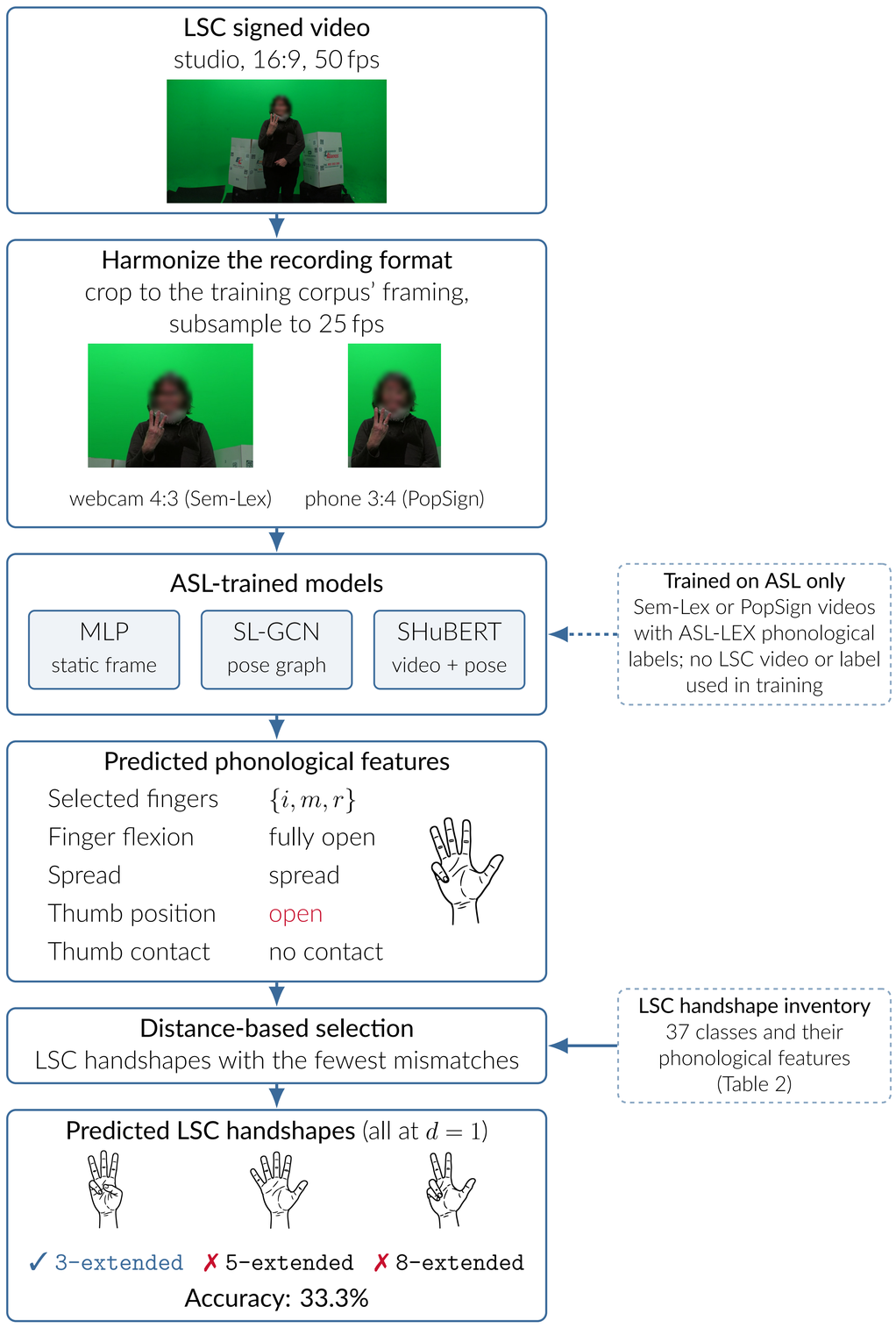


Figure 1. Zero-shot pipeline: the dashed boxes are the only language-specific inputs. Prediction example B from Table 3.

From phonological features to LSC handshapes

LSC-Benchmark handshape annotations follow the SignHub grammar [8]; we re-encode them with the five ASL-LEX phonological features:

LSC handshape	Selected fingers	Finger flexion	Spread	Thumb position	Thumb contact
`3-extended`	$\{i, m, r\}$	fully open	spread	closed	no contact
`8-extended`	$\{i, m\}$	fully open	spread	open	no contact
`5-extended`	$\{i, m, r, p\}$	fully open	spread	open	no contact
`B-extended`	$\{i, m, r, p\}$	fully open	unspread	open	no contact
`8-flat_closed`	$\{i, m\}$	flat	unspread	open	contact
⋮	*… and 32 more (37 benchmark handshapes, 43 in the grammar)*				

Table 2. LSC handshape inventory with ASL-LEX phonological features. Selected fingers: i index, m middle, r ring, p pinky.

Each phonological feature is predicted independently; the LSC handshapes with the fewest mismatched features (distance d) are the predicted handshapes:

Examples	Phonological features	Distance d
Prediction A: in the LSC inventory	$\{i, m, r\}$, fully open, spread, closed, no contact	–
→ `3-extended` (ground truth)	$\{i, m, r\}$, fully open, spread, closed, no contact	0
Predicted LSC handshapes {`3-extended`} ⇒ *accuracy*: 100%		
Prediction B: not in the LSC inventory	$\{i, m, r\}$, fully open, spread, open, no contact	–
→ `3-extended` (ground truth)	$\{i, m, r\}$, fully open, spread, closed, no contact	1
→ `5-extended`	$\{i, m, r, p\}$, fully open, spread, open, no contact	1
→ `8-extended`	$\{i, m\}$, fully open, spread, open, no contact	1
5 further handshapes at $d = 2$ (excluded)		2
Predicted LSC handshapes {`3-extended`, `5-extended`, `8-extended`} ⇒ *accuracy*: 33.3%		

Table 3. Examples of LSC handshape prediction. Blue: true handshape; red: misprediction; underlined: mismatch.

Results

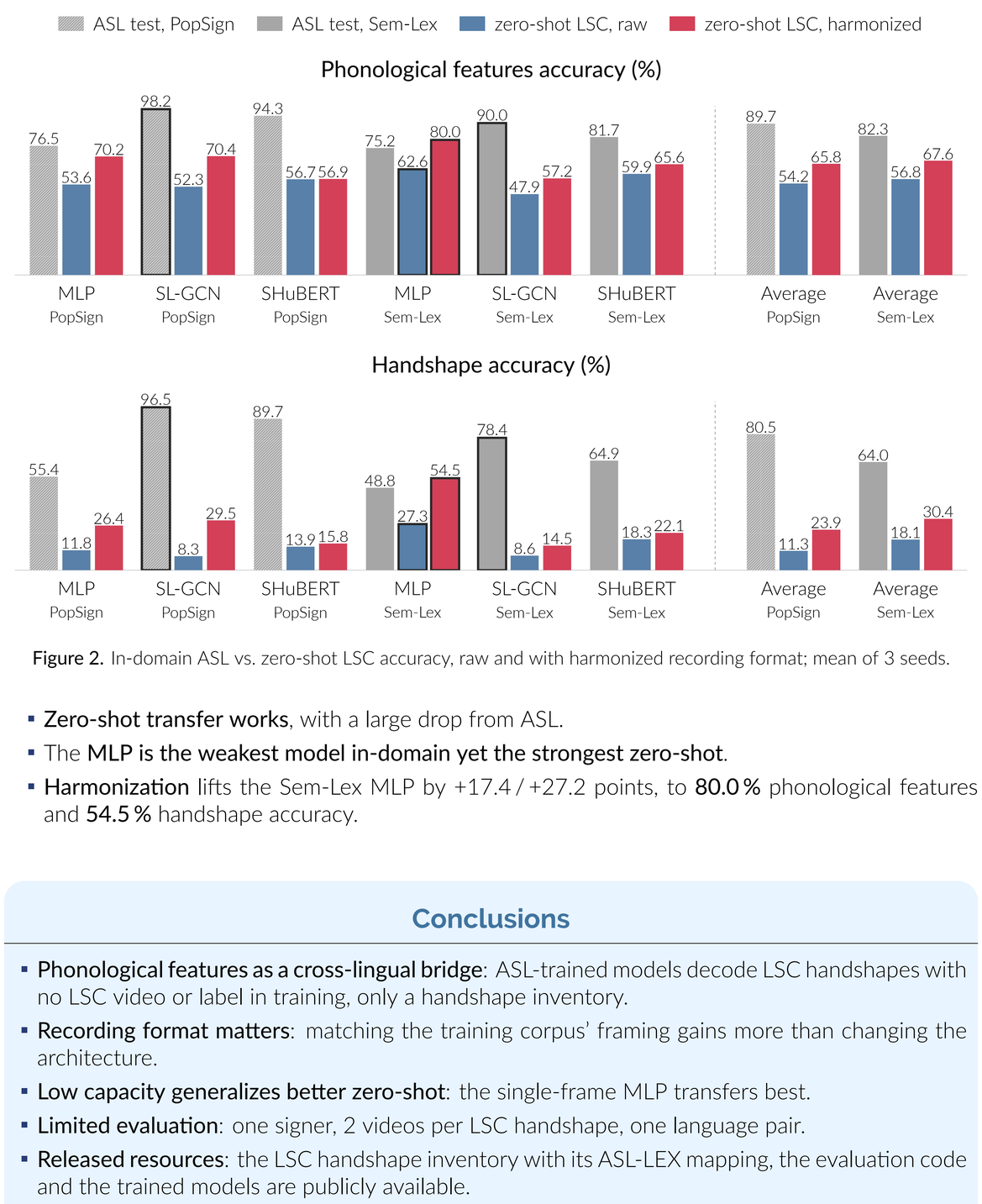


Figure 2. In-domain ASL vs. zero-shot LSC accuracy, raw and with harmonized recording format; mean of 3 seeds.

- **Zero-shot transfer works**, with a large drop from ASL.
- The **MLP is the weakest model in-domain yet the strongest zero-shot**.
- **Harmonization** lifts the Sem-Lex MLP by +17.4 / +27.2 points, to **80.0 %** phonological features and **54.5 %** handshape accuracy.

Conclusions

- **Phonological features as a cross-lingual bridge**: ASL-trained models decode LSC handshapes with no LSC video or label in training, only a handshape inventory.
- **Recording format matters**: matching the training corpus' framing gains more than changing the architecture.
- **Low capacity generalizes better zero-shot**: the single-frame MLP transfers best.
- **Limited evaluation**: one signer, 2 videos per LSC handshape, one language pair.
- **Released resources**: the LSC handshape inventory with its ASL-LEX mapping, the evaluation code and the trained models are publicly available.

Acknowledgements

This work is part of Project PID2025-173709NB-I00 (MuReLA), funded by MICIU/AEI/10.13039/501100011033. It was also supported by the Maria de Maeztu Units of Excellence Programme (CEX2021-001195-M, funded by MICIU/AEI/10.13039/501100011033), the PULSAR project (PID2025-173459NB-C22, financed by MICIU/AEI/10.13039/501100011033 and FEDER/UE), and VISTASL (2024-1-EL01-KA220-HED-000257847), funded by the EU's Erasmus+ programme. M. G. was supported by the Joan Oró predoctoral programme (AGAUR-FI, 2024 FI-1 01038) of the Department of Research and Universities of the Generalitat de Catalunya, co-funded by the European Social Fund Plus. G. H. acknowledges support from the Serra Húnter Programme (Generalitat de Catalunya). We acknowledge the EuroHPC Joint Undertaking for access to MareNostrum5 (BSC) and Leonardo (CINECA).

References

[1] D. Brentari. *A Prosodic Model of Sign Language Phonology*. MIT Press, 1998.
[2] L. Kezar et al. The Sem-Lex Benchmark: Modeling ASL Signs and Their Phonemes. In *ASSETS*, 2023.
[3] T. Starner et al. PopSign ASL v1.0: An Isolated ASL Dataset Collected via Smartphones. In *NeurIPS Datasets and Benchmarks*, 2023.
[4] Z. S. Sehyr et al. The ASL-LEX 2.0 Project: A Database of Lexical and Phonological Properties for 2,723 Signs in ASL. *JDSDE*, 2021.
[5] A. Carbo and E. Nalisnick. Improving Handshape Representations for SLP: A Graph Neural Network Approach. In *EMNLP*, 2025.
[6] S. Jiang et al. Skeleton Aware Multi-Modal Sign Language Recognition. In *CVPR Workshops*, 2021.
[7] S. Gueuwou et al. SHuBERT: Self-Supervised Sign Language Representation Learning via Multi-Stream Cluster Prediction. In *ACL*, 2025.
[8] A. Navarrete-González. Phonology: 1. Sublexical Structure. In *A Grammar of Catalan Sign Language (LSC)*. SIGN-HUB, 2020.