DINO Eats Geometry: Self-Supervised Sign Representations that Absorb Articulated Hand Geometry

Junyi Hu Zhewen He Haomian Huang Zhenhua Li
Shuaihang Yuan Zhifei Li Yi Fang*
New York University Abu Dhabi
jh10472@nyu.edu

Abstract

Handshape and orientation are geometric components of sign language, yet many sign encoders learn from appearance alone. We introduce SIGNDEG (*DINO Eats Geometry*), a temporal sign encoder that uses articulated hand geometry as a pre-training target and requires only video at inference. Frozen DINOV3 features of tracked hands and faces feed per-stream temporal Transformers. Geometric latent prediction (GLP) supervises an intermediate block with WiLoR latents, MANO articulation, and 3D joints, including at masked frames. Geometric relation transfer (GRT) matches cross-video frame-similarity distributions at the final block. Both objectives are learned jointly with temporal self-distillation; WiLoR and the geometry heads are then discarded. Under the SHUBERT public-data evaluation protocol, SIGNDEG reaches 17.9/24.5 BLEU-4 on How2Sign/OpenASL, compared with 16.2/23.2 for SHUBERT. Rank-1 adaptation improves ASL Citizen and Sem-Lex recognition by 0.02 Recall@1 each, and phonological classification improves across 16 attributes. Removing GRT while retaining GLP reduces How2Sign BLEU-4 by 1.4 and ASL Citizen Recall@1 by 0.034. The results support transferring articulated structure into an appearance encoder without running a hand reconstruction model at deployment.

1 Introduction

Gloss-free sign language translation (SLT) maps signing video to spoken-language text without intermediate gloss annotations (Lin et al., 2023; Wong et al., 2024; Li et al., 2025). Scarce parallel data motivate self-supervised sign encoders (Hu et al., 2021; Zhao et al., 2024; Rust et al., 2024; Gueuwou et al., 2025b; Gan et al., 2026). Sign phonology decomposes a sign into handshape, palm orientation, body-relative location, and movement, together with non-manual markers (Stokoe, 1960; Brentari, 1998). Handshape and orientation describe an articulated state; movement organises those states over time. Learning a useful sign representation therefore requires both fine articulatory structure and temporal context.

Appearance encoders supply strong visual features but do not explicitly supervise this articulated state. Whole-frame DINO adaptation (Gan et al., 2026) and temporal self-distillation over tracked appearance crops learn from image features whose similarity can reflect viewpoint or signer appearance as well as handshape. Systems that take pose as input (Hu et al., 2023; Li et al., 2025) make articulation explicit, but require a pose estimator at inference. A complementary approach is to use geometry as a learning target: the encoder can acquire articulatory structure during pre-training while retaining an appearance-only inference pathway.

We propose SIGNDEG (Figure 1). A frozen DINOV3 maps tracked hand and face crops to appearance features, and a temporal Transformer encodes each stream. During pre-training, a frozen hand reconstruction model supplies geometric targets to the hand streams. GLP predicts its latent, MANO articulation, and 3D joints from an intermediate block; masked frames require prediction from neighbouring motion. GRT aligns the final block's cross-video similarity distributions with those of the geometric targets. This division places per-frame articulation within the temporal encoder and geometric relations at its output. Both losses are active from the first update; the hand model is removed at inference. Using an auxiliary foundation-model space for training has proved useful in representation alignment and geometric generation (Yu et al., 2025; Jang et al., 2026); here the target is the articulated hand, and the downstream objective is sign representation learning.

Our contributions are: (i) a sign encoder that distils articulated hand geometry into temporal

* Corresponding author.

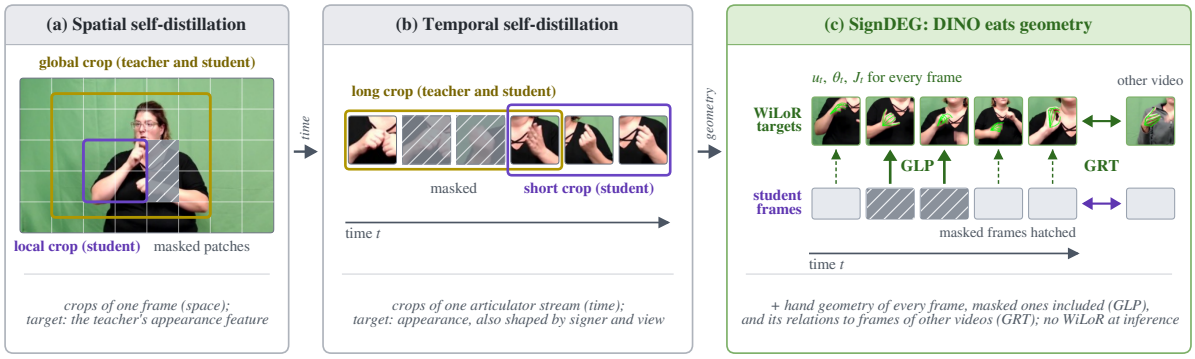


Figure 1: From spatial to temporal self-distillation to SIGNDEG. *(a)* Image DINO: a student matches an EMA teacher across global and local crops of one frame. *(b)* Temporal self-distillation plays the same game on long and short temporal crops of one tracked articulator stream, with some frames masked; in both, the target is the teacher's own appearance feature. *(c)* During pre-training, SIGNDEG also asks the student to predict, at every frame including masked ones, the geometry that a frozen hand reconstruction model (WiLoR) assigns to the crop (GLP), and to transfer geometric relations between frames of different videos (GRT). WiLoR is removed at inference.

appearance features without geometry estimation at inference; (ii) masked geometric prediction at intermediate depth and cross-video geometric relation transfer at the final block, with reliability-weighted targets; and (iii) evaluation across translation, recognition, fingerspelling, and phonological categories. Frozen features improve How2Sign translation over raw DINOv3 crops; rank-1 adaptation improves ASL Citizen recognition over SHUBERT. Matched geometry controls and label-based retrieval identify the structure contributed by the geometric objectives.

2 Related Work

Self-supervised sign representations. Sign SSL has used hand-pose reconstruction (Hu et al., 2021, 2023), masked video modelling (Zhao et al., 2024; Rust et al., 2024), sign–text alignment (Jiang et al., 2024), and prediction of clustered units over hand, face, and body streams (Gueuwou et al., 2025b). Multi-stream recognition (Chen et al., 2022; Yin and Read, 2020) and graph-based translation (Gan et al., 2025) model complementary spatial and temporal cues. The CVPR SIGNDINO (Gan et al., 2026) adapts frame-level DINO using whole-frame and articulator-focused views. SIGNDEG instead applies self-distillation along the time axis of tracked articulator streams and supervises this pathway with fixed hand-geometry targets. Geometry is a pre-training target rather than a downstream input; source signing videos need no gloss, translation, or geometry annotations for this stage.

Foundation-model features as learning targets. Representation alignment (Yu et al., 2025) and geometric latent diffusion (Jang et al., 2026; Lin et al., 2026; Wang et al., 2025) use foundation-model features to expose structure useful for another learning objective. DINO Eats CLIP (He et al., 2026) uses CLIP-guided virtual features for open-set 3D retrieval. Relational distillation transfers pairwise structure between feature spaces (Park et al., 2019). WiLoR and HaMeR recover MANO articulation and 3D joints from hand crops (Romero et al., 2017; Potamias et al., 2025; Pavlakos et al., 2024). Our contribution combines per-frame articulated targets with masked temporal prediction and cross-video relation transfer for sign language. WiLoR contributes a supervised geometric prior learned outside the signing corpus; pre-training the sign encoder itself uses no new manual annotations.

3 SIGNDEG

SIGNDEG uses a temporal encoder per articulator stream, with objectives assigned to different depths (Figure 2). Its input is appearance: a frozen DINOv3 embeds tracked crops (§3.1). Self-distillation between a student and an EMA teacher trains its output to describe how the stream evolves (§3.2). During pre-training, a frozen hand foundation model supplies per-frame geometric targets (§3.3); an intermediate block learns to predict them at every frame (§3.4), and the last block learns cross-video frame relations in geometric space (§3.5). Inference retains only the appearance pathway and the EMA teacher (§3.6). Throughout,

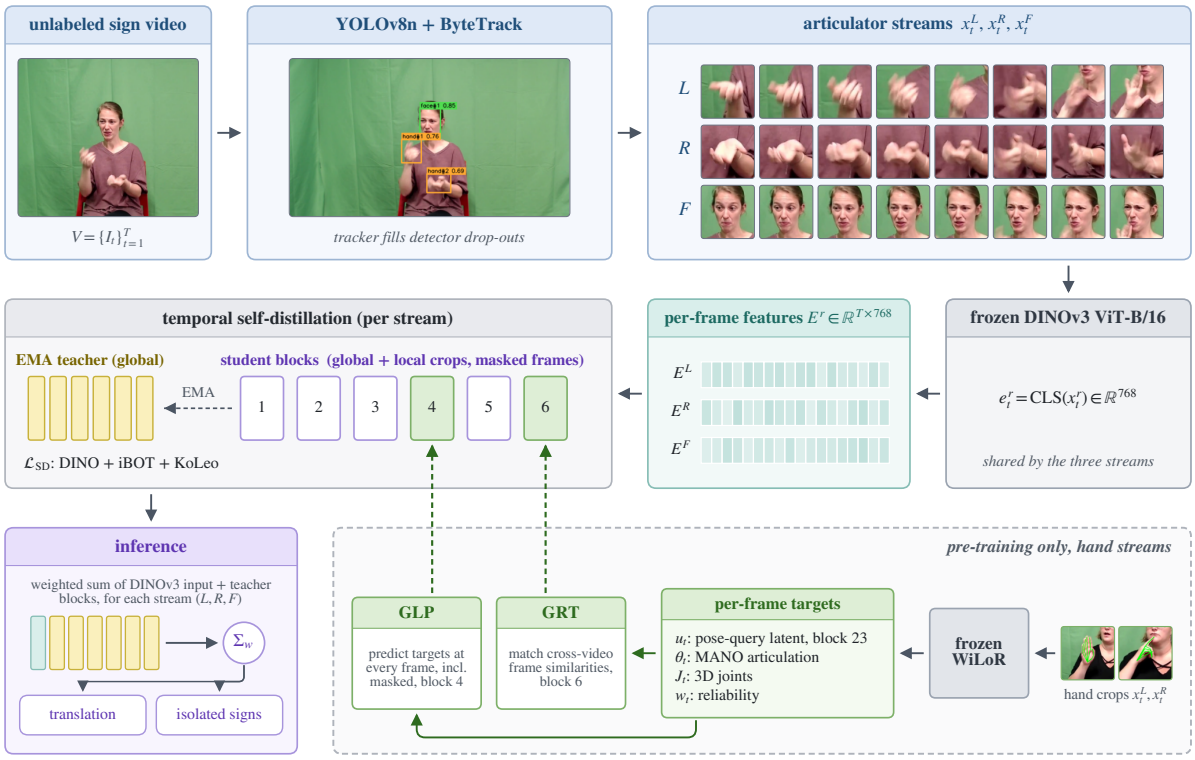


Figure 2: Overview of SIGNDEG. *Top:* a detector and tracker extract left-hand, right-hand, and face crop streams from each video. *Middle (right to left):* a frozen DINOv3 embeds every crop, and a temporal student and its EMA teacher are trained per stream by temporal self-distillation. *Bottom right:* during pre-training only, a frozen hand reconstruction model supplies per-frame geometric targets for the hand streams; the student predicts them from block 4, including at masked frames (GLP), and aligns its cross-video frame similarities at block 6 with geometric similarities (GRT). *Bottom left:* at inference, the EMA teachers run on DINOv3 features alone.

an *articulator crop* is the image region of one hand or the face in one frame, and a *temporal crop* is a contiguous run of frames of one stream.

3.1 Articulator streams

The hands and face occupy small regions of the frame and move independently, so each is tracked as a separate stream. A hand and face detector (YOLOv8n) runs on every frame, and ByteTrack (Zhang et al., 2022) associates hand boxes with two persistent identities and fills short detector gaps with its motion model. Each video yields three streams $r \in \{\mathrm{L}, \mathrm{R}, \mathrm{F}\}$ (left hand, right hand, face); x_t^r is the square articulator crop of stream r at frame t, and $m_t^r \in \{0, 1\}$ marks whether the tracker produced a box. A frozen DINOv3 ViT-B/16 (Siméoni et al., 2025) maps each crop to its class token:

$$e_t^r = \mathrm{CLS}\big(x_t^r\big) \in \mathbb{R}^{768}. \qquad (1)$$

These frame embeddings are cached once per video. Each stream has its own temporal Transformer, trained independently; we drop the index r below. Frame rate, crop sizes, and all other hyperparameters are listed in Appendix A.

3.2 Temporal-axis self-distillation

The output of a stream encoder should describe how the articulator evolves. We train it by temporal-axis self-distillation, which applies DINO's agreement between global and local views to the time axis (Figure 3). The student f_ϕ projects the sequence (e_t) to width D, prepends a class token, adds sinusoidal encodings of the frame index, and applies L pre-norm Transformer blocks; padded and undetected frames are excluded from attention. We write $z_t^{(l)}$ for the token of frame t after block l. The teacher $f_{\bar{\phi}}$ is the exponential moving average (EMA) of the student.

Each training sample draws two global and several local temporal crops, whose lengths are sampled from a long and a short range. The teacher sees the global crops. The student sees all crops, and in its global crops a random subset $\mathcal{M}$ of frames has its input replaced by a learned mask embedding. A projection head maps class tokens

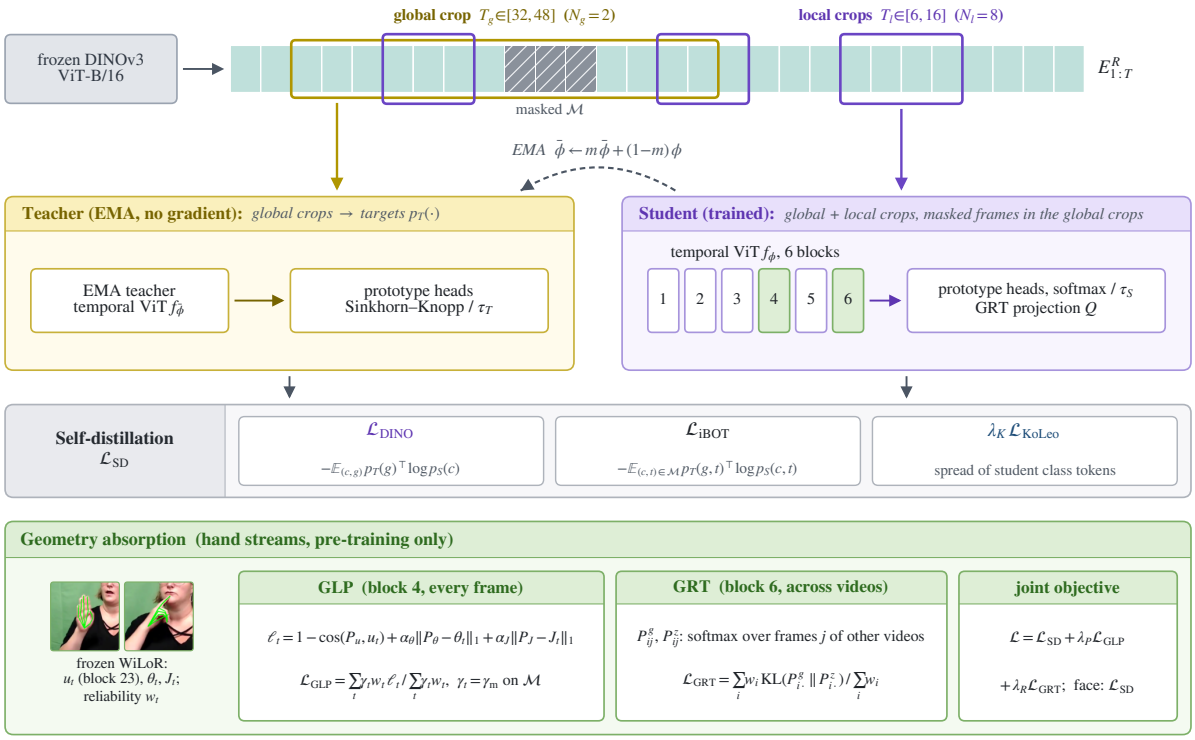


Figure 3: Training of one hand stream. Long (global) and short (local) temporal crops are drawn from the cached DINOv3 features; the EMA teacher sees the global crops, and the student sees all crops with some frames of its global crops masked. Self-distillation combines the DINO, iBOT, and KoLeo terms. GLP reads student frame tokens at block 4 and GRT reads them at block 6, using targets from frozen WiLoR with reliability weighting applied to each loss; both losses are added from the first step. The face stream uses self-distillation alone.

and frame tokens to K prototype scores. Writing $p_T(g)$ for the teacher's prototype distribution of global crop g, normalised over the batch by Sinkhorn–Knopp iterations (Caron et al., 2020), and $p_S(c)$ for the student's softmax distribution of crop c, the self-distillation objective combines a crop-level cross-entropy, its frame-level counterpart at masked frames (iBOT; Zhou et al., 2022), and the KoLeo entropy regulariser on student class tokens (Sablayrolles et al., 2019):

$$\mathcal{L}_{\mathrm{DINO}} = -\mathbb{E}_{(c,g)}\left[p_T(g)^\top \log p_S(c)\right], \quad (2)$$

$$\mathcal{L}_{\mathrm{iBOT}} = -\mathbb{E}_{(c,t)\in\mathcal{M}}\left[p_T(g,t)^\top \log p_S(c,t)\right], \quad (3)$$

$$\mathcal{L}_{\mathrm{SD}} = \mathcal{L}_{\mathrm{DINO}} + \mathcal{L}_{\mathrm{iBOT}} + \lambda_K \mathcal{L}_{\mathrm{KoLeo}}, \quad (4)$$

where the first expectation runs over pairs of a student crop and a different teacher global crop, and the second over masked frames of the student global crops paired with the same frames in the teacher.

3.3 Geometric targets from a hand foundation model

A frozen WiLoR (Potamias et al., 2025) supplies three targets for each available hand crop; left hands are mirrored into its right-hand convention. The geometric latent $u_t \in \mathbb{R}^{1280}$ is the mean of its 16 MANO pose-query tokens after backbone block k, standardised channel-wise using source-training statistics. The articulation $\theta_t \in \mathbb{R}^{15\times 6}$ represents the 15 local finger-joint rotations of MANO (Romero et al., 2017) in 6D (Zhou et al., 2019), excluding global wrist rotation. Its coordinate definition is independent of camera viewpoint. Wrist-relative joints $J_t \in \mathbb{R}^{21\times 3}$ remain in the camera frame and retain orientation. These are WiLoR estimates, not motion-capture ground truth. Extraction and scale conventions are specified in Appendix B.

The latent should encode articulation while remaining predictable from appearance. We select k=23 by the product of two ridge-probe scores: articulation predicted from the latent, and the latent predicted from DINOv3 crop features. Table 1 reports the candidate-layer comparison: block 23 has the highest product, balancing articulatory information and learnability from appearance. Ridge fitting uses 10,000 source-training frames and evaluation uses 2,000 video-disjoint source-development

Block k	$R^2(u \to \theta)$	$R^2(e \to u)$	Product
7	0.54	0.72	0.389
15	0.68	0.65	0.442
23	0.79	0.59	0.466
31	0.84	0.49	0.412

Table 1: Latent-layer selection on source-development videos. Block 23 balances articulatory predictability and learnability from appearance.

frames, sampled equally from both hands. Appendix B gives the fitting details. Selection uses source-development videos; downstream test labels are not used.

Reliability. Per-frame regressors fail on blurred and interacting hands, and such failures appear as abrupt jumps in θ. We weight frame t by

$$w_t = m_t\, o_t\, s_t\, \rho_t \exp\!\big(-\kappa\, \overline{|\theta_{t-1} - 2\theta_t + \theta_{t+1}|}\big), \tag{5}$$

where o_t indicates that WiLoR produced valid estimates for the frame and both neighbours, s_t is the detector score, and ρ_t is 0.5 for a tracker-filled box and 1 otherwise. The bar averages coordinates of the second temporal difference of the 6D rotations. At video boundaries or missing-neighbour frames, o_t=0. Neighbours are adjacent frames of the downsampled source video, independently of temporal-crop boundaries. This is a temporal smoothness proxy for target reliability: both reconstruction jitter and genuine rapid changes can reduce it. Low-weight frames still contribute to self-distillation; a uniform-weight control is reported in §4.5.

3.4 Geometric latent prediction

GLP trains the student's intermediate block l_g to encode the articulatory state of each frame. Lightweight heads P_u, P_θ, P_J ($384 \to 256 \to d$ MLPs with GELU, $d \in \{1280, 90, 63\}$) read the frame tokens of the student's global crops after this block, giving a per-frame loss ℓ_t that is averaged with weights:

$$\begin{aligned}\ell_t = {} & 1 - \cos\big(P_u(z_t^{(l_g)}), u_t\big) \\ & + \alpha_\theta \,\|P_\theta(z_t^{(l_g)}) - \theta_t\|_1 \\ & + \alpha_J \,\|P_J(z_t^{(l_g)}) - J_t\|_1,\end{aligned} \tag{6}$$

$$\mathcal{L}_{\mathrm{GLP}} = \frac{\sum_t \gamma_t w_t\, \ell_t}{\max(\sum_t \gamma_t w_t, \epsilon)}, \tag{7}$$

where $\|\cdot\|_1$ averages coordinates, γ_t=γ_{m}>1 for masked frames and 1 otherwise, and $\epsilon > 0$ prevents an empty valid-target set from producing a division by zero. Empty sets contribute zero geometry loss. At a masked frame, the student infers the target from the visible temporal context. We supervise block l_g=4 of L=6, leaving two blocks above it for temporal self-distillation and GRT; Table 6 compares this placement with final-block supervision.

3.5 Geometric relation transfer

GLP supplies a target per frame. GRT additionally transfers relations between different videos at the last block. From the batch's global crops, we deduplicate frames by (video, stream, frame index), then uniformly sample up to N=1024 valid frames with $w_i > 0$ across the distributed batch. Student descriptors are $\hat{q}_i = Q(z_i^{(L)})/\|Q(z_i^{(L)})\|$; geometric descriptors $\hat{g}_i$ concatenate separately ℓ_2-normalised u_i and θ_i and are normalised again. Let $\mathcal{N}(i)$ contain sampled frames from videos other than i's. Anchors with no such candidates are omitted. With

$$P^g_{ij} = \frac{\exp(\hat{g}_i^\top \hat{g}_j/\tau_g)}{\sum_{j' \in \mathcal{N}(i)} \exp(\hat{g}_i^\top \hat{g}_{j'}/\tau_g)} \tag{8}$$

and P^z_{ij} defined in the same way from $\hat{q}$ with temperature τ_z, for $j \in \mathcal{N}(i)$, GRT minimises

$$\mathcal{L}_{\mathrm{GRT}} = \frac{\sum_i w_i\, \mathrm{KL}\big(P^g_{i\cdot} \,\|\, P^z_{i\cdot}\big)}{\max(\sum_i w_i, \epsilon)}. \tag{9}$$

Candidate distributions are unweighted; w_i scales the KL term of anchor i. Excluding the anchor video removes immediate temporal neighbours, making the objective transfer geometric relations beyond one trajectory. Video identity is available without signer labels; different videos can contain the same signer. GRT therefore imposes cross-video, rather than guaranteed cross-signer, supervision. Its relation descriptors emphasise articulation; the GLP joints retain camera-relative orientation. We evaluate cross-signer handshape grouping separately using human phonological labels (§4.4).

3.6 Objective and inference

The two hand streams are trained with

$$\mathcal{L} = \mathcal{L}_{\mathrm{SD}} + \lambda_P\, \mathcal{L}_{\mathrm{GLP}} + \lambda_R\, \mathcal{L}_{\mathrm{GRT}}, \tag{10}$$

and the face stream with $\mathcal{L}_{\mathrm{SD}}$ alone. All geometry targets are fixed and all terms are active from the first update. The inference encoders are the three EMA teachers; WiLoR and P_u, P_θ, P_J, Q are discarded. Thus geometry adds no inference modules to the same temporal appearance backbone. A

task-specific learned softmax mixes the teacher's L block outputs and its projected DINOv3 input (Yang et al., 2021); the three stream features are projected and concatenated for the downstream head. Offline feature caching accelerates training, while processing a new video still requires detection, tracking, and the image encoder.

4 Experiments

We follow SHuBERT's public-data evaluation suite for source-corpus accounting, target splits, task heads, metrics, and downstream schedules. Published baselines provide task-level context; the within-backbone controls in Table 6 isolate geometric supervision. The CVPR SignDINO (Gan et al., 2026) is a separate frame-level method. Appendix G gives the training protocol; Appendix B gives geometry diagnostics.

4.1 Pre-Training Setup

We pre-train SignDEG on the YouTube-ASL source split of SHuBERT (Uthus et al., 2023; Gueuwou et al., 2025b), in which clips intersecting the OpenASL evaluation set are replaced with non-overlapping ASL videos from YouTube-SL-25 (Tanzer and Zhang, 2025) (≈984 source hours). Each stream has a 10.8M-parameter temporal Transformer (L=6, D=384, 6 heads) on the shared frozen DINOv3 ViT-B/16. Temporal encoders and DINO/iBOT heads comprise approximately 47M trainable parameters; auxiliary geometry heads are accounted separately in Appendix B. All runs use 4×A100 80 GB GPUs; training schedules and compute are in Appendices G and H.3.

4.2 Sign Language Translation

The translation head replicates SHuBERT §4.2: the layer-weighted teacher features of each stream are normalised, projected, concatenated, and fed to a ByT5-Base decoder (Xue et al., 2022). Training has two stages (Table 12): a source stage on the YouTube-ASL split above and a target stage on How2Sign or OpenASL; FLEURS-ASL is scored zero-shot after the source stage. The *live* variant updates the temporal encoders at $1/10$ of the decoder learning rate, and the *frozen* variant trains only the fusion projection and ByT5. We report BLEU-4 (Papineni et al., 2002) via SacreBLEU (Post, 2018) and BLEURT-20 (Sellam et al., 2020).

Results. Live-fine-tuned SignDEG reaches *17.9*/*24.5* BLEU-4 on How2Sign/OpenASL (Table 2), *+1.7*/*+1.3* BLEU-4 and *+1.4*/*+1.1* BLEURT over SHuBERT, and *+2.43*/*+7.04* BLEU-4 over the published SignDINO results. With frozen encoders it reaches 16.8/23.6, exceeding the published live SHuBERT result. WiLoR and the geometry heads are absent at test time. SignDEG also exceeds Uni-Sign on OpenASL (24.5 vs. 23.1) and SHuBERT on zero-shot FLEURS-ASL (5.3 vs. 4.7). Reading only the last temporal layer instead of the layer-weighted sum lowers How2Sign to 15.8 (14.7 frozen; Table 19), indicating that multiple encoder depths contribute to decoding.

4.3 Isolated Sign Language Recognition

We evaluate ISLR on ASL Citizen (Desai et al., 2023), Sem-Lex (Kezar et al., 2023), and WLASL2000 (Li et al., 2020), the SHuBERT suite (MSASL is excluded because its test set overlaps YouTube-ASL). The head follows SHuBERT §4.3: a stream-fused, mask-aware temporal average followed by BatchNorm and a linear classifier. We report three regimes: frozen features with a linear head, rank-1 LoRA (Hu et al., 2022) on the temporal encoders (0.11M parameters), and full fine-tuning of the temporal encoders.

Fingerspelling detection. On ASL-STEM-Wiki (Yin et al., 2024), detection is per-frame binary classification with a two-layer Transformer over the fused features, followed by thresholding, median smoothing, and interval extraction; we report mean interval-IoU under the dataset's cross-validation protocol (Appendix G).

Results. With rank-1 LoRA (Table 3), SignDEG improves Recall@1 over SHuBERT by *+0.02* on both ASL Citizen (*0.67*) and Sem-Lex (*0.56*); full fine-tuning reaches *0.704* and *0.593*. On WLASL2000, LoRA gives *66.9*/*64.2* P-I/P-C, above Uni-Sign (*63.52*/*61.32*) and SHuBERT (*60.90*/*58.01*), and full fine-tuning gives *68.5*/*65.9*. Fingerspelling mIoU is *0.41* with LoRA (SHuBERT *0.40*) and *0.43* with full fine-tuning. Frozen features with a linear head reach *0.63*/*0.52* R@1 on ASL Citizen/Sem-Lex; the gains over SHuBERT are obtained with matched rank-1 adaptation.

4.4 Phonological and Articulatory Probing

We train 16 classifiers for ASL-Lex 2.0 categories following SHuBERT (Appendix E). Mean R@1 is 87.64/91.85 on Sem-Lex/ASL Citizen, compared

Method	SSL	PT hrs	How2Sign		OpenASL		FLEURS-ASL
			BLEU↑	BLEURT↑	BLEU↑	BLEURT↑	BLEU↑
SSVP-SLT (Rust et al., 2024)	✓	1054	15.5	49.6	—	—	—
YouTube-ASL fine-tune (Uthus et al., 2023)	×	984	12.4	46.6	—	—	—
YouTube-SL-25 (Tanzer and Zhang, 2025)	×	3207	15.4	47.9	—	—	4.4
SignMusketeers (Gueuwou et al., 2025a)	✓	984	14.3	—	—	—	—
SpaMo (Hwang et al., 2025)	×	—	10.11	42.23	—	—	—
Uni-Sign (Li et al., 2025)	×	984	14.9	49.4	23.1	60.4	—
OpenASL (Shi et al., 2022)	×	—	—	—	6.7	31.1	—
SHuBERT (Gueuwou et al., 2025b)	✓	984	16.2	49.9	23.2	60.6	4.7
SignDINO (Gan et al., 2026)	✓	—	15.47	—	17.46	—	—
SignDEG							
SignDEG frozen + ByT5 (frozen encoders)	✓	984	$16.8_{\pm 0.3}$	$50.4_{\pm 0.4}$	$23.6_{\pm 0.3}$	$60.9_{\pm 0.3}$	$4.9_{\pm 0.2}$
SignDEG live fine-tune (updated encoders)	✓	984	$\mathbf{17.9_{\pm 0.15}}$	$\mathbf{51.3_{\pm 0.4}}$	$\mathbf{24.5_{\pm 0.3}}$	$\mathbf{61.7_{\pm 0.3}}$	$\mathbf{5.3_{\pm 0.2}}$
SignDEG live, last-layer only (no weighted sum)	✓	984	$15.8_{\pm 0.3}$	$48.7_{\pm 0.4}$	$23.0_{\pm 0.3}$	$59.7_{\pm 0.4}$	$4.6_{\pm 0.2}$

Table 2: Sign language translation under the SHuBERT public-data protocol. PT hrs: source pre-training video hours. SignDEG is pre-trained on the OpenASL-decontaminated YouTube-ASL split, fine-tuned on How2Sign or OpenASL, and evaluated zero-shot on FLEURS-ASL; Frozen/live denote whether the temporal encoders are updated; this is the freezing convention of SHuBERT, not equivalence to its last-layer ablation numbers. SignDEG entries are mean ± std over 3 seeds; baselines are published results. Dashes indicate values not reported in the cited work.

Method	Trainable	ASL Citizen			Sem-Lex			WLASL2000		ASL-STEM-Wiki
		R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	P-I↑	P-C↑	mIoU↑
Baselines										
ST-GCN (Desai et al., 2023)	0.45M	0.60	0.82	0.88	—	—	—	—	—	—
SignCLIP (Jiang et al., 2024)	217M	0.60	0.84	0.89	0.30	0.48	0.55	—	—	—
I3D (Desai et al., 2023)	25M	0.63	0.86	0.91	—	—	—	—	—	—
SignBERT+ (Hu et al., 2023)	—	—	—	—	—	—	—	48.85	46.37	—
MSLU (Zhou et al., 2025)	—	—	—	—	—	—	—	56.29	53.29	—
Uni-Sign (Li et al., 2025)	580M	—	—	—	—	—	—	63.52	61.32	—
SHuBERT (rank-1 LoRA) (Gueuwou et al., 2025b)	0.17M	0.65	0.87	0.91	0.54	0.74	0.80	60.90	58.01	0.40
SignDEG										
SignDEG frozen + linear	linear	$0.63_{\pm .01}$	$0.86_{\pm .01}$	$0.91_{\pm .01}$	$0.52_{\pm .01}$	$0.72_{\pm .01}$	$0.79_{\pm .01}$	$60.1_{\pm 0.4}$	$57.5_{\pm 0.4}$	$0.39_{\pm .01}$
SignDEG rank-1 LoRA	0.11M	$0.67_{\pm .01}$	$0.88_{\pm .01}$	$0.92_{\pm .01}$	$0.56_{\pm .01}$	$0.76_{\pm .01}$	$0.82_{\pm .01}$	$66.9_{\pm 0.4}$	$64.2_{\pm 0.4}$	$0.41_{\pm .01}$
SignDEG full fine-tune	all	$\mathbf{0.704_{\pm .01}}$	$\mathbf{0.89_{\pm .01}}$	$\mathbf{0.93_{\pm .01}}$	$\mathbf{0.593_{\pm .01}}$	$\mathbf{0.78_{\pm .01}}$	$\mathbf{0.84_{\pm .01}}$	$\mathbf{68.5_{\pm 0.4}}$	$\mathbf{65.9_{\pm 0.4}}$	$\mathbf{0.43_{\pm .01}}$

Table 3: Isolated sign recognition (Recall@k; WLASL2000 per-instance/per-class top-1) and fingerspelling detection (ASL-STEM-Wiki mIoU) under the SHuBERT suite; Sem-Lex uses SHuBERT's full test set. SignDEG: mean ± std over 3 seeds. Dashes indicate values not reported in the cited work.

with 79.38/85.02 for SHuBERT. On Sem-Lex, Handshape improves from 62.93 to 71.16, Selected Fingers from 79.53 to 89.94, and Spread from 79.42 to 90.99. These labels are independent of WiLoR. A frozen ridge probe on How2Sign explains 56% of the variance in WiLoR-estimated articulation versus 44% for raw DINOv3 features, measuring transfer of the teacher's geometric estimates. A frozen cross-signer retrieval diagnostic additionally tests human-labelled handshape across different lexical signs; its results are in Table 4.

Frozen cross-signer handshape retrieval. We use ASL Citizen's official signer-disjoint test queries and training-gallery clips that map to a valid ASL-Lex 2.0 dominant-handshape label. No model is fine-tuned for this diagnostic. We average valid hand-stream frame features (block 4 for temporal models), concatenate the two hand streams, and ℓ_2-normalise. We rank gallery clips by cosine similarity, excluding the query signer and query lexical sign. P@5 is the fraction of the five retrieved clips with the same human-assigned handshape, averaged first within handshape and then across supported handshapes. We use all eligible queries and at most ten gallery clips per lexical sign, sampled deterministically with seed 0. This tests shared articulation across signers and lexical items; it does not test calibrated multi-view geometry.

Table 4 reports P@5 of 0.61 for full SignDEG, compared with 0.41 for raw DINOv3 and 0.49 for SD only. GLP alone reaches 0.55, while adding GRT further improves grouping across signers and lexical signs.

4.5 Ablations and Representation Diagnostics

Table 6 isolates GLP and GRT within the same architecture and training schedule. Table 7 retains temporal-context and tracking controls; Table 5 removes articulator streams. Masking, data scale,

Frozen representation	Handshape P@5	Articulation R^2
Raw DINOv3	0.41	0.44
SD only	0.49	0.48
SD + GLP (w/o GRT)	0.55	0.54
Full SIGNDEG	0.61	0.56

Table 4: Frozen geometry diagnostics. Handshape retrieval uses independent human labels and excluded-signer/excluded-lexeme candidates; articulation R^2 predicts WiLoR estimates on How2Sign.

Streams kept at fusion head	BLEU
LH + Face (no RH)	9.8
RH + Face (no LH)	11.2
LH + RH (no Face)	14.6
LH + RH + Face (default)	17.9

Table 5: Leave-one-stream-out at the fusion head, How2Sign test BLEU-4 (live fine-tuning; SSL unchanged).

layer pooling, backbone, decoder, and crop-length comparisons appear in Appendix H.5, and substitutions of the hand detector and the hand geometry model in Appendix H.5.

Stream contribution. Dropping the right-hand stream costs the most (*-8.1* BLEU from 17.9), followed by the left hand (*-6.7*) and the face (*-3.3*), in line with the predominantly right-handed signers of YouTube-ASL and How2Sign; the face adds mouthing and non-manual cues.

Geometry at fixed backbone and schedule. Removing GRT while retaining GLP lowers How2Sign from 17.9 to 16.5 BLEU and ASL Citizen from 0.704 to 0.670 R@1. Relative to SD only (15.6 BLEU, 0.642 R@1), GLP and GRT separately improve both tasks, and their combination reaches the highest scores. Moving GLP to block 6 gives 17.2 BLEU and 0.690 R@1, supporting intermediate supervision with temporal layers above it. All controls hold the image backbone, crops, encoder capacity, and downstream heads fixed.

Reliability-weighted articulation probes. Using the same valid How2Sign fitting frames, reliability-weighted ridge regression with Eq. (5) reaches R^2=0.56, compared with 0.53 under uniform weights. This offline control evaluates how target weighting affects probe fitting, with the SSL encoders fixed. Appendix B gives the fitting details.

Temporal context and tracking. Restricting the teacher to short windows lowers H2S BLEU/ASLC R@1 to 15.8/0.65; removing ByteTrack gives 16.4/0.61 (Table 7). The tracked streams and long temporal context contribute alongside geometry.

Objective / placement	H2S BLEU	ASLC R@1
SD only	15.6	0.642
SD + GLP (w/o GRT)	16.5	0.670
SD + GRT (w/o GLP)	16.4	0.665
SD + GLP + GRT (block 4)	17.9	0.704
SD + GLP + GRT (block 6)	17.2	0.690

Table 6: Matched geometry controls: identical 984-hour source split, 450K updates, and downstream schedules. H2S uses live SLT; ASLC uses full temporal fine-tuning.

Ablation	H2S BLEU	ASLC R@1
Full SIGNDEG	17.9	0.704
w/o GRT (λ_R=0)	16.5	0.67
Teacher short window only	15.8	0.65
YOLO only (no ByteTrack)	16.4	0.61

Table 7: Ablations of GRT (λ_R=0 in Eq. (10), GLP kept), teacher temporal context, and tracking.

Additional design-axis and protocol ablations. The frame-only mask gives the best BLEURT (Table 18); the randomised crop-length sampler beats fixed windows by *0.4–0.6* BLEU (Table 21); How2Sign BLEU spans *16.5–18.0* over five frozen DINO backbones (Table 22); and SIGNDEG beats SHUBERT by *+1.4–+1.7* BLEU across five decoders (Table 23). Going from 10% to 100% of the pre-training data adds *+3.3* BLEU.

Training and downstream diagnostics. Prototype-use entropy remains above 0.91; task-specific layer weights spread across more depths for SLT than for recognition and fingerspelling. Appendix H interprets these as prototype-occupancy and layer-use diagnostics, and Appendix D provides descriptive clusters.

5 Conclusion

We presented SIGNDEG, which transfers hand-model targets into an intermediate temporal layer and geometric frame relations into the final layer. The geometry model is used only for pre-training. Across ASL translation, recognition, fingerspelling, and phonological classification, the resulting representation improves over the published SHUBERT task results in matched adaptation regimes. The GRT removal control costs 1.4 BLEU-4 on How2Sign and 0.034 R@1 on ASL Citizen; matched geometry controls and phonological retrieval assess the additional roles of GLP, layer placement, and cross-video relations.

Limitations

SIGNDEG depends on hand and face crops: prolonged detector or tracker dropouts remove visual evidence. WiLoR provides estimated geometry and imports the biases and errors of its supervised training; local articulation coordinates do not guarantee view-invariant estimates. The temporal smoothness weight may also downweight rapid signing. GRT excludes same-video neighbours but does not require signer identity during pre-training. The human-labelled retrieval diagnostic tests cross-signer handshape grouping, while calibrated cross-view reconstruction is outside this evaluation. Experiments cover ASL and ASL-to-English, and the three-stream model does not explicitly encode upper-body pose. Privacy, consent, and Deaf-community participation remain important when extending to other languages or deploying these representations (Desai et al., 2024). New-video inference still includes the detector, tracker, and image backbone, and hand-target extraction adds pre-training cost.

References

Diane Brentari. 1998. *A Prosodic Model of Sign Language Phonology*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924.

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

Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard E. Ladner, Hal Daumé III, Alex X. Lu, Naomi Caselli, and Danielle Bragg. 2023. ASL citizen: A community-sourced dataset for advancing isolated sign language recognition. In *Advances in Neural Information Processing Systems*, volume 36, pages 76893–76907.

Aashaka Desai, Maartje De Meulder, Julie A. Hochgesang, Annemarie Kocab, and Alex X. Lu. 2024. Systemic biases in sign language AI research: A Deaf-Led call to reevaluate research agendas. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 54–65, Torino, Italia. ELRA and ICCL. ArXiv:2403.02563.

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.

Shiwei Gan, Xiao Liu, Yafeng Yin, Nan Liu, Kuizhuang Liu, Desibieer Tuerdaken, Zhiwei Jiang, Lei Xie, Sanglu Lu, and Hongkai Wen. 2026. Learning effective sign features without text for gloss-free sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9827–9836.

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Lei Xie, Sanglu Lu, and Hongkai Wen. 2025. MixSignGraph: A sign sequence is worth mixed graphs of nodes. In *Advances in Neural Information Processing Systems*, volume 38, pages 139371–139395.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, and Karen Livescu. 2025a. SignMusketeers: An efficient multi-stream approach for sign language translation at scale. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22506–22521. ArXiv:2406.06907.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, Karen Livescu, and Alexander H. Liu. 2025b. 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*, pages 28792–28810. ArXiv:2411.16765.

Xinwei He, Yansong Zheng, Qianru Han, Zhichuan Wang, Yuxuan Cai, Yang Zhou, Jingbo Xia, Yulong Wang, Jinhai Xiang, and Xiang Bai. 2026. DINO eats CLIP: Adapting beyond knowns for open-set 3D object retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 34704–34713.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. 2023. SignBERT+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11221–11239.

Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. 2021. SignBERT: Pre-training of hand-model-aware representation for sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11087–11096.

Eui Jun Hwang, Sukmin Cho, Junmyeong Lee, and Jong C. Park. 2025. An efficient gloss-free sign language translation using spatial configurations and motion dynamics with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3901–3920, Albuquerque, New Mexico. Association for Computational Linguistics.

Wooseok Jang, Seonghu Jeon, Jisang Han, Jinhyeok Choi, Minkyung Kwon, Seungryong Kim, Saining Xie, and Sainan Liu. 2026. Repurposing geometric foundation models for multi-view diffusion. In *Proceedings of the European Conference on Computer Vision (ECCV)*. ArXiv:2603.22275.

Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. SignCLIP: Connecting text and sign language by contrastive learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9171–9193. ArXiv:2407.01264.

Lee Kezar, Jesse Thomason, Naomi Caselli, Zed Sevcikova Sehyr, and Elana Pontecorvo. 2023. The Sem-Lex benchmark: Modeling ASL signs and their phonemes. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–10. ArXiv:2310.00196.

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*, pages 1448–1458.

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)*.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. 2026. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations (ICLR)*.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. Gloss-free end-to-end sign language translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12904–12916.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 7 others. 2024. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.

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.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3962–3971.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. 2024. Reconstructing hands in 3D with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9826–9836.

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.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. 2025. WiLoR: End-to-end 3D hand localization and reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12242–12254.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 36(6):245:1–245:17.

Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgöz, and Jean Maillard. 2024. Towards privacy-aware sign language translation at scale. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8624–8641.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. 2019. Spreading vectors for similarity search. In *International Conference on Learning Representations (ICLR)*.

Zed Sevcikova Sehyr, Naomi Caselli, Ariel M. Cohen-Goldberg, and Karen Emmorey. 2021. The ASL-LEX 2.0 project: A database of lexical and phonological properties for 2,723 signs in american sign language. *The Journal of Deaf Studies and Deaf Education*, 26(2):263–277.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

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.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, and 7 others. 2025. DINOv3. *Preprint*, arXiv:2508.10104.

William C. Stokoe. 1960. *Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf*. Number 8 in Studies in Linguistics: Occasional Papers. Department of Anthropology and Linguistics, University of Buffalo, Buffalo, New York.

Garrett Tanzer. 2025. FLEURS-ASL: Including american sign language in massively multilingual multitask evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6167–6191.

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)*.

Dave Uthus, Garrett Tanzer, and Manfred Georg. 2023. YouTube-ASL: A large-scale, open-domain american sign language-english parallel corpus. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks)*, volume 36, pages 29029–29047.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. 2025. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5294–5306.

Ryan Wong, Necati Cihan Camgöz, and Richard Bowden. 2024. Sign2GPT: Leveraging large language models for gloss-free sign language translation. In *International Conference on Learning Representations (ICLR)*.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

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.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2021. SUPERB: Speech processing universal PERformance benchmark. In *Proc. Interspeech*, pages 1194–1198.

Kayo Yin and Jesse Read. 2020. Better sign language translation with STMC-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989.

Kayo Yin, Chinmay Singh, Fyodor O. Minakov, Vanessa Milan, Hal Daumé III, Cyril Zhang, Alex Xijie Lu, and Danielle Bragg. 2024. ASL STEM Wiki: Dataset and benchmark for interpreting STEM articles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14474–14490. ArXiv:2411.05783.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. 2025. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations (ICLR)*.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. 2022. ByteTrack: Multi-object tracking by associating every detection box. In *Computer Vision – ECCV 2022*, volume 13682 of *Lecture Notes in Computer Science*, pages 1–21. Springer.

Weichao Zhao, Hezhen Hu, Wengang Zhou, Yunyao Mao, Min Wang, and Houqiang Li. 2024. MASA: Motion-aware masked autoencoder with semantic alignment for sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):10793–10804.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. 2022. iBOT: Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations (ICLR)*.

Wengang Zhou, Weichao Zhao, Hezhen Hu, Zecheng Li, and Houqiang Li. 2025. Scaling up multimodal pre-training for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):11753–11767. Earlier preprint: arXiv:2408.08544.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5738–5746.

A SIGNDEG Pre-training Hyperparameters

Parameter	Value
Frame rate	12.5 fps
Crop size / padding	224^2; hands $1.4\times$, face $1.2\times$
Temporal encoder	L=6, D=384, 6 heads
Global / local crops	$2 \times [32, 48]$ / $8 \times [6, 16]$ frames
Masked fraction	$\mathcal{U}[0.25, 0.50]$
Prototypes K	8192
Teacher / student temperature	0.04→0.07, then 0.07 / 0.1
λ_K	0.1
λ_P, λ_R	1.0, 1.0
α_θ, α_J, γ_{m}	1.0, 1.0, 2.0
l_g, k (probed: 7, 15, 23, 31)	4, 23
GRT: N, τ_g, τ_z	1024, 0.05, 0.1
Reliability κ, ρ	3.0, 0.5
Downstream projection per stream	256
Optimiser	AdamW, lr 5×10^{-4} then 2×10^{-4}, wd $0.04 \rightarrow 0.4$
EMA momentum	$0.994 \rightarrow 1.0$, then $0.999 \rightarrow 1.0$
Batch / updates	32 videos / 400K + 50K (Table 11)

Table 8: SIGNDEG pre-training hyperparameters (all streams; geometry terms on the hand streams only).

B Geometric Target and Diagnostic Protocols

Extraction conventions. WiLoR's 32-block, 1280-dimensional ViT includes 16 MANO pose-query tokens. Local rotations exclude the wrist's global rotation; wrist-relative joints retain the camera axes. We use the released WiLoR checkpoint and native crop normalisation, read backbone blocks with zero-based indices 7/15/23/31, and extract the refined MANO articulation and joints. We mirror left-hand input horizontally into the right-hand convention, keep local finger rotations in that canonical convention, and reflect wrist-relative joints back to original camera axes with $S = \mathrm{diag}(-1, 1, 1)$. We express J in metres and fit per-channel latent statistics on source-training frames only ($\epsilon = 10^{-6}$).

Implementation specification. We use the existing custom YOLOv8n hand/face checkpoint. Its weight hash, category mapping, and detector-training manifest will be released with the configurations. For a frontal signer, we initialise anatomical left/right tracks from image-right/image-left respectively and retain their ByteTrack IDs through crossing motions. Cases with ambiguous initial side assignment are flagged in the validity mask. Tracker-filled boxes retain the last detector score; geometry neighbours are adjacent frames of the original downsampled video, not temporal-crop boundaries. For GRT, we deduplicate overlapping global-crop frames by (video, stream, frame index), then uniformly sample up to 1024 valid frames across the distributed batch; we remove same-video candidates and omit anchors with no remaining candidates. Candidate distributions are unweighted; Eq. (9) weights anchors.

Each GLP head is $384 \rightarrow 256 \rightarrow d$, with GELU and $d \in \{1280, 90, 63\}$; Q is a linear $384 \rightarrow 128$ projection. Separate DINO/iBOT heads use $384 \rightarrow 512 \rightarrow 256 \rightarrow 8192$, with GELU in hidden layers and a normalised bottleneck before prototypes. Geometry heads add approximately 1.4M student parameters across the two hand streams (approximately 48.4M total trained parameters with the 47M temporal/prototype core); they are not EMA inference modules. Downstream inference uses full attention over each benchmark clip, with padded/invalid frames masked and omitted from temporal averaging.

Latent selection and ridge probes. We sample 10,000 valid source-training frames and 2,000 source-development frames, with video-disjoint splits and equal sampling from the two hands. We fit ridge regressions $u^{(k)} \rightarrow \theta$ and $e \rightarrow u^{(k)}$ after standardising predictors on the fitting split. We select the regularisation coefficient from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ by a video-disjoint inner split. We average coordinate-wise test R^2 and multiply the two development scores to select k. Table 1 reports the scores; no downstream test videos enter layer selection. The How2Sign articulation probes yield 44%/56% explained variance, with the same fitting procedure applied on How2Sign train/validation/test respectively.

The frozen cross-signer handshape-retrieval protocol and results are reported in §4.4 and Table 4.

Reliability and cost. We stratify retained target weight by motion-speed quartile to check suppression of rapid valid signing. WiLoR target extraction adds approximately 160 A100 GPU-hours for both hands; combined pre-training plus appearance caching and geometry extraction is then approximately 1300 GPU-hours. These costs are reported in GPU-hours; the SHUBERT figure uses different hardware.

C Default Hyperparameter Card

Table 9 is the high-level configuration card; the SSL optimisation segments (Table 11) and the

source/target translation training schedule (Table 12) are reported in Appendix G.

Hyperparameter	Default value
Backbone	frozen DINOv3 ViT-B/16
Streams	LH, RH, face (3 streams, independent SSL)
Temporal encoder	L=6, D=384, 6 heads, MLP ratio 4
Crops (contiguous)	$T_g \sim \mathcal{U}\{32-48\}$, $T_l \sim \mathcal{U}\{6-16\}$; N_g=2, N_l=8
Temperatures	τ_T : $0.04 \rightarrow 0.07$, $\tau_S = 0.1$; Table 11
EMA momentum	cosine, $0.994 \rightarrow 1.0$ then $0.999 \rightarrow 1.0$ (Table 11)
Prototype K	8192
Geometry weights λ_P, λ_R	1.0, 1.0
KoLeo weight	0.1
Optimiser	AdamW, cosine decay, bf16
Detector + tracker	YOLOv8n + ByteTrack; gaps $\leq$ 3 frames filled
Downstream SLT	ByT5-Base, beam = 5, $L_{\max}$=384, label smoothing 0.2
Downstream ISLR	BN + linear; rank-1 LoRA; 125 epochs, batch 128
Downstream FS	2-layer Transformer + frame head; threshold 0.5

Table 9: Default hyperparameter card. Values summarise the default configuration used for the main experiments and ablations.

EMA momentum schedule under variable-length crops. EMA momentum follows the per-update cosine schedules in Table 11. Each video supplies $N_g T_g + N_l T_l \in [112, 224]$ student frame tokens, with expectation 168. DINO averages crop pairs, iBOT averages valid masked positions, and KoLeo averages class tokens. These normalisations avoid scaling the loss directly with crop length; they do not imply identical gradient variance across crop samplers. Geometry losses are normalised by their valid target weights.

D Qualitative Cluster Analysis of SIGNDEG Features

To examine what the SIGNDEG encoder captures, we run a fully zero-shot qualitative analysis on held-out frames, mirroring the cluster-visualisation protocol of SHUBERT Appendix B (Figures 5–7 there) (Gueuwou et al., 2025b). After the SSL training described in §3.2 — no signing gloss or translation labels, and with OpenASL clips that overlap with the YouTube-ASL pre-training source already decontaminated at the SSL stage (§4.1) — we apply the frozen SIGNDEG pipeline to a fresh sample of *previously unseen* How2Sign and OpenASL videos: YOLOv8n + ByteTrack extracts synchronised face / left-hand / right-hand crops, each crop is embedded through the frozen DINOv3 ViT-B/16 (§3.1), and the per-stream CLS sequences are processed by the trained temporal EMA encoders before their output vectors are L2-normalised. We then run k-means (K=250 for the face stream, K=200 for each hand stream) on these embeddings and select clusters by a farthest-first traversal in centroid space, filtered to clusters whose members span a sufficiently diverse set of source videos (so single-signer / single-shot clusters are excluded). For each chosen cluster we render the 10 nearest-to-centroid crops. The procedure is intended to be *descriptive*: we do not use the cluster labels in any downstream evaluation. Figures 4–6 show the result. Notably, the face stream surfaces clusters defined by *expression and pose* (hand-on-face mouthings, side-profile, neutral closed mouth) that recur across multiple identities — analogous to SHUBERT cluster 141 — and each hand stream surfaces *handshape* clusters (pointing / spread-fingers / thumb-up / two-hand contact) that recur across many backgrounds and skin tones, qualitative evidence consistent with sub-lexical grouping, though not proof of signer or background invariance. Anatomical hand labels follow the signer's perspective: "left hand" (Figure 5) corresponds to image-right under our convention, since signers face the camera.

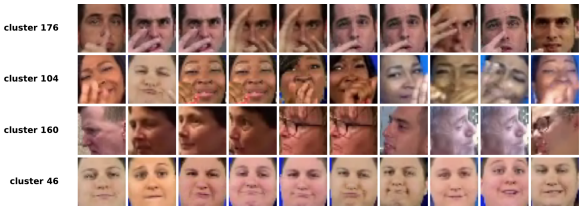


Figure 4: Sample *face* clusters from SIGNDEG features on held-out How2Sign and OpenASL frames. Four k-means clusters (K=250); each row shows 10 nearest-to-centroid members. Cluster 176 captures "hand-on-face mouthings" across multiple identities; cluster 104 captures hand-near-mouth gestures with varied skin tones; cluster 160 surfaces side-profile / looking-down poses; cluster 46 collects neutral frontal views with a closed mouth.

E Phonological Feature Recognition

Motivation and protocol. Following SHUBERT Appendix B/C, we include a phonological-feature recognition benchmark in addition to task-level evaluation. This probe asks whether a learned video representation preserves the sub-lexical structure from which ASL signs are formed, rather than only supporting a whole-sign class label or an English translation. We adopt this benchmark because it is a direct diagnostic of whether the model has encoded deeper sign-language information: lo-

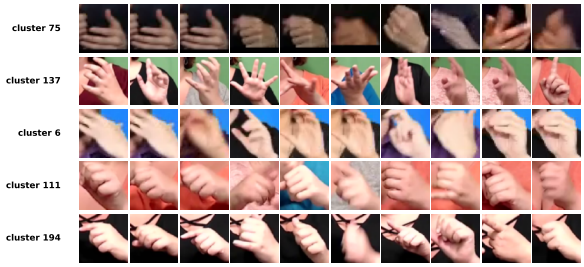


Figure 5: Sample *left-hand* clusters from SIGNDEG features on held-out How2Sign and OpenASL frames. Five k-means clusters (K=200). Cluster 75 captures thumb-up / closed-fist handshapes; cluster 137 captures the spread-fingers ("5") handshape across multiple backgrounds; clusters 6 and 111 capture mid-motion signing gestures; cluster 194 captures two-hand contact poses.

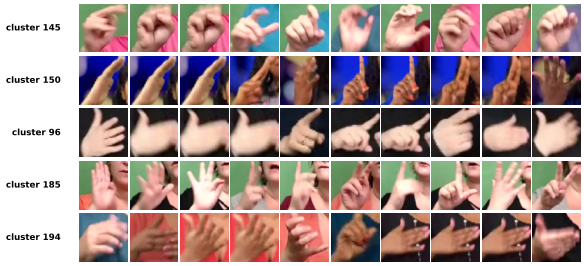


Figure 6: Sample *right-hand* clusters from SIGNDEG features on held-out How2Sign and OpenASL frames. Five k-means clusters (K=200). Cluster 145 captures pointing / extended-index handshapes across multiple backgrounds; cluster 150 captures a related forward-pointing pose on dark backgrounds; cluster 96 captures the thumb-up / closed-fist family; cluster 185 captures open-palm / spread-fingers gestures; cluster 194 captures motion-blurred active gesturing.

cation, handshape, selected fingers, contact, and motion are not incidental visual attributes, but linguistically meaningful components of ASL signs. If a representation supports accurate prediction of these properties across datasets, it provides stronger evidence that the SSL model has captured sign-internal structure and category information.

We use Sem-Lex and ASL Citizen with the ASL-Lex 2.0 feature inventory (Kezar et al., 2023; Desai et al., 2023; Sehyr et al., 2021). Each masked temporal average feeds 16 linear heads. The SHUBERT rows are published full-fine-tuning results, reported as per-feature top-1 accuracy. For the SIGNDEG rows, we fully fine-tune the temporal encoders and heads with the image backbone frozen, using Adam with learning rate 10^{-4} for both, no weight decay, batch 128, and at most 125 epochs. We use seed 0 for this single-run attribute table, select checkpoints by the mean validation accuracy across supported attributes, and mask missing labels separately for each head's cross-entropy. We preserve the public train/validation/test video splits and the same ASL-Lex label mapping as the comparator. The separately frozen retrieval diagnostic in §4.4 does not use these adapted encoders.

Phonological feature	Method	Sem-Lex R@1	ASL Citizen R@1
Major Location	SHUBERT	84.77	90.22
	SIGNDEG	**94.43**	**94.96**
Minor Location	SHUBERT	71.30	80.00
	SIGNDEG	**76.83**	**85.79**
Second Minor Location	SHUBERT	73.28	81.18
	SIGNDEG	**82.34**	**90.73**
Contact	SHUBERT	86.84	91.57
	SIGNDEG	**98.93**	**96.94**
Thumb Contact	SHUBERT	84.74	87.52
	SIGNDEG	**92.55**	**92.16**
Sign Type	SHUBERT	84.64	91.54
	SIGNDEG	**90.72**	**99.50**
Repeated Movement	SHUBERT	82.65	89.93
	SIGNDEG	**87.00**	**96.21**
Path Movement	SHUBERT	72.75	79.42
	SIGNDEG	**81.12**	**87.72**
Wrist Twist	SHUBERT	90.58	93.00
	SIGNDEG	**97.11**	**99.50**
Selected Fingers	SHUBERT	79.53	83.44
	SIGNDEG	**89.94**	**87.67**
Thumb Position	SHUBERT	86.04	88.19
	SIGNDEG	**97.28**	**98.76**
Flexion	SHUBERT	72.64	77.73
	SIGNDEG	**78.74**	**82.83**
Spread	SHUBERT	79.42	84.80
	SIGNDEG	**90.99**	**91.89**
Spread Change	SHUBERT	81.60	86.58
	SIGNDEG	**86.44**	**91.75**
Nondominant Handshape	SHUBERT	76.32	84.32
	SIGNDEG	**86.60**	**93.63**
Handshape	SHUBERT	62.93	70.80
	SIGNDEG	**71.16**	**79.51**
Average	SHUBERT	79.38	85.02
	SIGNDEG	**87.64**	**91.85**

Table 10: Phonological feature recognition on Sem-Lex and ASL Citizen. The table follows the SHUBERT Appendix B layout; R@1 is top-1 accuracy for each feature-specific classifier. For every feature we report the SHUBERT baseline above and SIGNDEG below; SIGNDEG rows are bolded to mark the matched-protocol gain.

Phonology results. Averaged over the 16 phonological features, SIGNDEG reaches *87.64* R@1 on Sem-Lex and *91.85* R@1 on ASL Citizen, compared with SHUBERT's *79.38* and *85.02*, a matched-protocol gain of *+8.26* and *+6.83* R@1 respectively. SIGNDEG improves every feature on both datasets, including temporally articu-

lated features — Path Movement (*+8.37/+8.30*), Wrist Twist (*+6.53/+6.50*), Repeated Movement (*+4.35/+6.28*) — and fine handshape categories — Selected Fingers (*+10.41/+4.23*), Flexion (*+6.10/+5.10*), Handshape (*+8.23/+8.71*). On the categorical features that were already strong for SHUBERT, SIGNDEG also closes most of the remaining headroom: Major Location reaches *94.43/94.96* R@1 (*+9.66/+4.74*) and Sign Type reaches *90.72/99.50* (*+6.08/+7.96*). The consistent gains across the 16 measured attributes indicate that the full SIGNDEG representation supports more accurate prediction of these sub-lexical categories than the SHUBERT features. This result does not, on its own, distinguish the effects of temporal self-distillation, GLP, and GRT.

To the best of our knowledge, SHUBERT was the first work to report phonological-feature recognition accuracies on ASL Citizen under this setting, and the Sem-Lex numbers are not directly comparable to the original Sem-Lex report because SHUBERT evaluates on the entire public test set. We therefore follow SHUBERT and treat this benchmark as a reusable diagnostic for future sign representation learning work. Here, this benchmark is useful because SIGNDEG replaces discrete hidden-unit prediction with temporal self-distillation plus geometric supervision; the probe tests whether this continuous temporal objective still preserves the linguistic categories encoded by ASL-Lex. The full feature vocabulary used by the 16 classification heads is listed in Appendix K.

F Cross-Sentence t-SNE of Translation-Trained Features

Figure 7 shows a t-SNE of SIGNDEG features after SLT training. It visualises post-SLT features from a 2400-word sample spanning How2Sign and the three ISLR datasets, with perplexity 30 and 1000 iterations. It is a descriptive illustration, not a geometric diagnostic and not evidence used in model selection or held-out evaluation.

The noun and verb panels illustrate lexical neighbourhoods after translation supervision. Their projected distances are descriptive; the quantitative geometric evaluation uses the probes in Appendix B and signer-separated retrieval in §4.4.

G Reproducibility Checklist

This appendix records the training and evaluation protocol. Released artefacts will include per-stream checkpoints, benchmark features, and run configurations. Published baseline values are distinguished from SIGNDEG task results in Tables 2–3; matched geometry controls are reported in Table 6. Geometry ablation names in this manuscript refer to Eq. (10).

Hyperparameter	First 400K updates	Final 50K updates
Steps	400K	50K
Batch (videos / GPU)	8	8
Optimiser	AdamW	AdamW
Base LR	5×10^{-4}	2×10^{-4}
LR schedule	5% warmup + cosine	1% warmup + cosine
Weight decay	$0.04 \rightarrow 0.4$	$0.04 \rightarrow 0.4$
EMA momentum	$0.994 \rightarrow 1.0$	$0.999 \rightarrow 1.0$
Teacher τ_T	$0.04 \rightarrow 0.07$	0.07
Student τ_S	0.1	0.1
Multi-temporal crops	N_g=2, N_l=8	same
T_g/T_l ranges (contiguous)	$[32, 48]/[6, 16]$	same
iBOT mask ratio	25–50%	same
Loss weights	Eq. (10)	Eq. (10)
Precision	bf16 autocast	bf16 autocast

Table 11: SIGNDEG optimisation settings recorded over the initial 400K and final 50K updates. Both segments use the full loss of Eq. (10).

	Source stage	Target stage
Data	decontaminated YouTube-ASL	How2Sign / OpenASL train
Steps	250K	50K
Optimiser	AdamW (β_1=0.9, β_2=0.99)	AdamW
LR ByT5 + head	5×10^{-4}	1×10^{-4}
LR SIGNDEG (live)	5×10^{-5}	1×10^{-5}
Warmup	10K cosine	5K cosine
Effective batch	64 utterances	64 utterances
Weight decay	0.1	0.1
Label smoothing	0.2	0.2
Decoding	beam = 5, $L_{\max}$=384	beam = 5
Hardware	4 × A100 80GB	4 × A100 80GB

Table 12: Translation training hyperparameters. The source stage follows SHUBERT §4.2; the target stage fine-tunes on the original How2Sign/OpenASL training split and evaluates on the corresponding validation/test splits.

Source split and decontamination. For self-supervised pre-training and source-stage SLT training, we follow the SHUBERT public-data protocol. The source corpus is YouTube-ASL (Uthus et al., 2023). Clips that intersect the OpenASL evaluation set (Shi et al., 2022) are removed to avoid test contamination, and the removed duration is replaced with non-overlapping ASL videos from YouTube-SL-25 (Tanzer and Zhang, 2025). This keeps the public source corpus at approximately 984 source hours. We also do not report MSASL because its test set overlaps with YouTube-ASL, following the same rationale as SHUBERT.

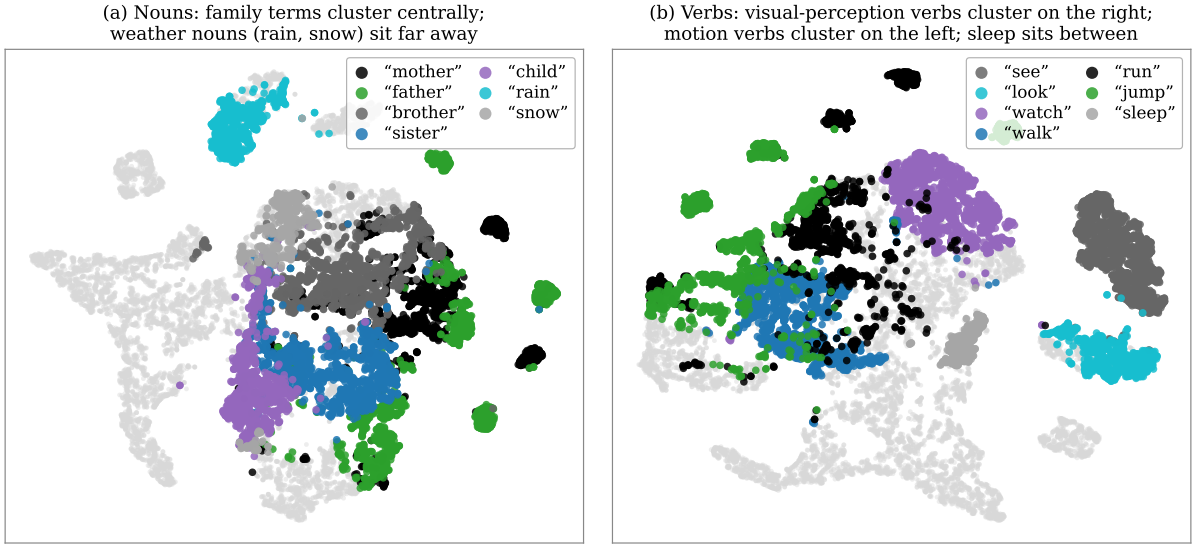


Figure 7: t-SNE of SIGNDEG features after SLT training. Noun and verb neighbourhoods are descriptive; this figure is not a geometry diagnostic.

Data and preprocessing. For source pre-training, each YouTube-ASL video is passed through YOLOv8n + ByteTrack to obtain synchronised left-hand, right-hand, and face crop streams. For target-stage fine-tuning, How2Sign and OpenASL videos are processed once from the original training split, matching the main SHUBERT-style protocol. Validation and test videos are also processed once. Each stream carries a binary validity mask indicating detector failures. Short detector dropouts are filled by ByteTrack's motion model; longer failures are excluded from SSL losses through the validity mask. Crops are square, enlarged $1.4\times$ (hands) and $1.2\times$ (face), and resized to 224×224 for the frozen DINOv3 encoder. In result tables, pre-training hours are always reported as unique source-video hours, matching SHUBERT's accounting.

Embedding cache and SSL training. Each crop stream is encoded once by the frozen DINOv3 ViT-B/16 into a float16 tensor $E^{(r)} \in \mathbb{R}^{T\times 768}$. All hand-stream training updates jointly optimise temporal self-distillation, GLP and GRT (Eq. (10)); face-stream training uses temporal self-distillation only. Optimisation has an initial block of 400K updates and a final block of 50K updates (Table 11), with GLP and GRT active in both. Global and local temporal crops and masked frames follow the settings of Appendix A and Eq. (7). The frozen hand-model targets are computed for the two hand streams during pre-training and discarded at inference.

Translation. The effective batch is 2 utterances/GPU $\times$ 4 GPUs $\times$ 8 accumulation steps $=$ 64 utterances. Translation follows SHUBERT's source-to-target adaptation protocol. The source stage trains the fusion projection and ByT5-Base translation model on the decontaminated YouTube-ASL split for 250K steps. The target stage fine-tunes on the original training split of How2Sign (Duarte et al., 2021) or OpenASL (Shi et al., 2022) for 50K steps. FLEURS-ASL (Tanzer, 2025) is evaluated zero-shot, since it has no training split in this protocol. Decoding uses beam search with width 5 and maximum output length 384. We report BLEU-4 (Papineni et al., 2002) computed with SacreBLEU (Post, 2018) and BLEURT (Sellam et al., 2020) computed with the BLEURT-20 checkpoint, following SHUBERT.

ISLR, fingerspelling, and phonology. ISLR evaluates ASL Citizen, Sem-Lex, and WLASL2000. The stream-fused representation is time-averaged with the attention mask, passed through BatchNorm and a linear classifier, and adapted either as frozen features, rank-1 LoRA, or full fine-tuning as described in §4.3. We report Recall@1/5/10 for ASL Citizen and Sem-Lex, and per-instance/per-class top-1 accuracy for WLASL2000. Fingerspelling detection on ASL-STEM-Wiki is trained as per-frame binary classification and evaluated by mean interval-IoU

under the dataset cross-validation protocol. The phonological-feature probe in Appendix E uses 16 simultaneous feature-specific heads and reports Recall@1 for each ASL-Lex feature.

Compute, software, and seeds. SSL uses bf16 and AdamW on 4×A100 80GB, training streams sequentially: 400K initial updates and 50K final updates, with all losses active in both segments. Table 15 reports the corresponding 840 temporal-training GPU-hours and approximately 300 GPU-hours for the appearance cache. WiLoR extraction is separately accounted in Appendix B. Software is PyTorch 2.6, CUDA 12.4, Transformers, SacreBLEU, and BLEURT-20. Three-seed summaries are identified explicitly in table captions; single-valued ablations use seed 0. Run configurations will record the seed identifiers and whether pre-training is shared. The matched geometry controls use SSL seed 0, downstream seeds 0/1/2, and validation-only checkpoint selection; Table 6 reports downstream means.

Responsible use. Sign-language video contains identifiable signer information. We use the released research corpora under their terms and do not redistribute raw video. Face crops retain non-manual linguistic cues; this is not an anonymisation procedure. Privacy-aware deployment requires a separate data-handling decision informed by prior sign-language research (Rust et al., 2024; Desai et al., 2024).

H Training and Downstream Diagnostics

This appendix reports prototype occupancy during SSL and task-specific use of encoder depth. The downstream task and phonological evaluations measure representation utility; the statistics here describe the training distributions.

H.1 Prototype-use entropy

Protocol. Let p_{bt} be a teacher frame's Sinkhorn-normalised assignment over $K{=}8192$ prototypes. We report $H(\bar{p})/\log K$, where $\bar{p}$ averages assignments over 1024 held-out frames. This is marginal prototype-use entropy: high values indicate broad occupancy. Balanced Sinkhorn assignments and even identical uniform outputs can have high marginal entropy, so this statistic alone is not a test of feature non-collapse.

Observed trajectory. Table 13 shows marginal entropy decreasing from 1.000 at initialisation to about 0.92 for hands and 0.94 for faces. The teacher continues to distribute assignment mass broadly. Differences between streams describe occupancy and do not establish a count of linguistic categories.

SSL checkpoint	LH	RH	Face
init	1.000	1.000	1.000
100K updates	0.962	0.958	0.971
200K updates	0.945	0.941	0.957
300K updates	0.933	0.929	0.945
400K updates	0.927	0.922	0.940
450K updates	0.921	0.917	0.935

Table 13: Marginal prototype-use entropy $H(\bar{p})/\log K$ on 1024 held-out frames. Values above 0.91 indicate broad occupancy; balanced assignments alone do not establish non-collapse.

H.2 Layer-weighted-sum head weights

Protocol. For each downstream task and each stream, the fusion head learns a softmax distribution over the $L{+}1{=}7$ SIGNDEGencoder layers ($L{=}0$: projected DINOv3 input; $L{=}1..6$: transformer blocks), as described in §4.2 and §4.3. We report the converged softmax weights after the target-stage / downstream training completes, averaged over three random seeds. Each row of Table 14 is therefore a probability distribution and sums to 1.0.

Observed patterns. The hand-stream weights peak at block 5 (0.21–0.27); recognition and fingerspelling place more combined weight on blocks 5 and 6 than translation does. Face weights are flatter. Because mixing coefficients depend on feature scale and are learned during downstream training, they describe how the head combines layers rather than proving a uniquely localised linguistic mechanism.

Task	Stream	$L{=}0$	$L{=}1$	$L{=}2$	$L{=}3$	$L{=}4$	$L{=}5$	$L{=}6$
SLT (H2S)	LH	0.06	0.09	0.12	0.16	0.18	0.21	0.18
	RH	0.05	0.08	0.11	0.15	0.19	0.22	0.20
	Face	0.07	0.11	0.14	0.17	0.19	0.18	0.14
ISLR (ASLC)	LH	0.04	0.07	0.10	0.14	0.18	0.24	0.23
	RH	0.04	0.07	0.10	0.14	0.18	0.24	0.23
	Face	0.09	0.12	0.15	0.17	0.17	0.16	0.14
FS (StemWiki)	LH	0.03	0.05	0.08	0.13	0.20	0.27	0.24
	RH	0.03	0.05	0.08	0.13	0.20	0.27	0.24
	Face	0.11	0.14	0.16	0.17	0.16	0.14	0.12

Table 14: Converged layer-weighted-sum softmax weights per task and per stream. Each row is a probability distribution over the seven encoder layers ($L{=}0$: projected DINOv3 input, $L{=}1..6$: transformer blocks) and sums to 1.0. Values are averaged over three random seeds of the target-stage / downstream training.

H.3 Pre-Training Compute Budget

Table 15 reports the wall-clock and GPU-hour budget of the SIGNDEG SSL pipeline on the $4 \times$ A100 80 GB hardware used for every main run in this paper. The three per-stream temporal Transformers are trained sequentially: each stream consumes the full 4-GPU pool for the duration of its optimisation schedule, then the next stream begins. With the frozen-DINOv3 embedding cache enabled (§4.1), per-step throughput is dominated by the temporal Transformer rather than by the image-backbone forward pass, which is what allows the full upstream pipeline to fit inside this budget. The embedding-cache row records the one-off preprocessing cost referenced in §4.1 and is amortised across all SSL ablations and downstream tasks. The final row reports the upstream SSL cost quoted by SHUBERT (Gueuwou et al., 2025b) as a reference point.

Stage	Per-stream GPU-h	Per-stream wall-clock	Total GPU-h
First 400K updates	240	≈60 h	720
Final 50K updates	40	≈10 h	120
SSL subtotal (3 streams)	280	≈70 h	840
Embedding cache (one-off)	—	—	≈300
SHUBERT upstream (ref.)	—	—	1344

Table 15: SIGNDEG pre-training compute budget on $4 \times$ A100 80 GB. Streams are trained sequentially, so the wall-clock for the full three-stream upstream is approximately $3\times$ the per-stream wall-clock plus negligible scheduling overhead. The embedding-cache row is the one-off preprocessing cost discussed in §4.1. The SHUBERT row reports the upstream-only figure of 1344 GPU-hours quoted by Gueuwou et al. (2025b) ($8 \times$ A6000 $\times$ 7 days).

H.4 SSL Data-Scaling Curve

Table 16 reports the two-point endpoints of the SSL data-scaling scan at 10% and 100% of the decontaminated YouTube-ASL source corpus. Table 17 extends this scan to a five-point pre-training-hours sweep along the same corpus, evaluated under the same frozen-features setup as Table 16 (SIGNDEGencoders frozen and the same downstream ByT5-Base head). For every row, the temporal sampler (Table 11), embedding cache, downstream source-stage + target-stage recipe (Table 12), and SSL optimisation schedule are held fixed at the defaults; only the size of the source subset used for SSL pre-training differs. Intermediate subsets are drawn as a video-disjoint random sample of the full source-video pool to keep the same speaker and topic distribution at every scale.

Pre-train hours	BLEU-1	BLEU	BLEURT
98 (10%)	33.6	13.5	46.1
984 (100%)	37.9	16.8	50.4

Table 16: Pre-training data scale (SHUBERT Tab. 5 analogue), evaluated on the *How2Sign test* split with the SIGNDEGencoders frozen and the same downstream head as Table 18. Two-point scan on 10% vs 100% of the decontaminated YouTube-ASL source corpus; PT hours follow the unique source-corpus accounting used in Table 2. Table 17 extends this to a five-point sweep.

Pre-train hours	H2S BLEU	H2S BLEURT
98 (10%)	13.5	46.1
246 (25%)	14.8	47.9
492 (50%)	16.0	49.4
738 (75%)	16.5	50.0
984 (100%)	16.8	50.4

Table 17: SIGNDEG SSL data-scaling curve. Frozen-features H2S BLEU and BLEURT as a function of the source-corpus pre-training budget. The 10% and 100% rows reproduce the corresponding rows of Table 16. The curve is monotonic and not yet flat at 100% of the public corpus, consistent with the data-scaling regime reported by SHUBERT (Gueuwou et al., 2025b); increasing hours from 98 to 984 yields 3.3 BLEU, with smaller gains at the upper end.

H.5 Additional Ablations

This appendix gives the full tables for the design-axis and SHUBERT-protocol ablations summarised in §4.5: masking strategy (Table 18), layer pooling (Table 19), frozen vs. live fine-tuning (Table 20), teacher crop-length randomisation (Table 21), visual backbone (Table 22), translation decoder (Table 23), and hand detector and hand geometry model (Table 24). The two-point and five-point data-scale results are in Tables 16–17 above. All rows reuse the common training protocol of Appendix G and are evaluated on the *How2Sign test* split unless noted; BLEU denotes corpus BLEU-4.

Layer pooling. Table 19 confirms that the SIGNDEGtemporal encoder makes a substantial contribution to translation: feeding the raw frozen DINOv3 per-frame embeddings to ByT5 without the SIGNDEGencoders drops to *9.4* BLEU/*41.2* BLEURT, a *-7.4* BLEU gap from the weighted-sum default. The last-layer-only variant (*14.7* BLEU/*47.8* BLEURT) recovers most but not all of the gap, motivating the learnable softmax

Strategy	BLEU-1	BLEU	BLEURT
Channel masking	35.2	14.4	47.6
Time masking	36.1	14.9	48.1
Random span masking	37.0	15.7	48.9
Frame mask only (SIGNDEG)	37.9	16.8	50.4
Frame + stream mask	37.3	16.3	49.7

Table 18: Masking-pattern controls within SIGNDEG, with frozen encoders and the same downstream head. Pattern names follow SHUBERT's masking comparison; the cells are SIGNDEG results, not published SHUBERT values. Frame-only masking gives the highest BLEURT.

over all $L{+}1$ encoder layers as the default pooling strategy.

Layer of SIGNDEG	BLEU-1	BLEU	BLEURT
None (raw DINOV3)	24.3	9.4	41.2
Last layer	34.6	14.7	47.8
Weighted sum (default)	37.9	16.8	50.4

Table 19: Layer pooling (SHUBERT Tab. 6 analogue), with the SIGNDEGencoders frozen and the same downstream head as Table 18. The "None" row feeds the raw DINOV3 per-frame embeddings to ByT5 without the SIGNDEGencoders.

Frozen vs. live fine-tune. The two rows of Table 20 reproduce the gap reported in the main Table 2: *16.8* BLEU frozen vs. *17.9* BLEU live (BLEURT *50.4→51.3*, BLEU-1 *37.9→39.4*). The *+1.1* BLEU gain from live fine-tuning is small relative to the cost of updating the temporal encoder; the frozen variant is therefore the recommended operating point when downstream compute is constrained, and live fine-tuning is the recommended operating point for the highest score in this comparison.

Fine-tune SIGNDEG?	BLEU-1	BLEU	BLEURT
× frozen	37.9	16.8	50.4
✓ live fine-tune	39.4	17.9	51.3

Table 20: Frozen vs. fine-tune (SHUBERT Tab. 7 analogue). This isolates whether downstream supervision benefits from updating the temporal SIGNDEGencoder beyond the frozen-feature setting; the two rows reproduce the corresponding "frozen" and "live fine-tune" rows of Table 2.

Teacher crop-length randomisation. The current default sampler draws global crops from $\mathcal{U}\{32,\dots,48\}$ and local crops from $\mathcal{U}\{6,\dots,16\}$, independently on each update (§3.2). At $N_g{=}2$ and $N_l{=}8$, this amounts to an expected $2\cdot 40+8\cdot 11=$ 168 sampled frames per update. Table 21 compares three fixed-window configurations $(T_l, T_g) \in \{(16,64),(20,80),(24,72)\}$, which use respectively 256, 320, and 336 sampled frames per update. The recorded BLEU-4 values are 17.3, 17.5, and 17.4, versus 17.9 for the randomised default. Because the frame budgets differ, the comparison is descriptive.

T_l / T_g sampler	H2S BLEU	Δ
Fixed 16 / 64	$17.3_{\pm 0.15}$	-0.6
Fixed 20 / 80	$17.5_{\pm 0.15}$	-0.4
Fixed 24 / 72	$17.4_{\pm 0.15}$	-0.5
Random $\mathcal{U}\{6{-}16\}$ / $\mathcal{U}\{32{-}48\}$ (default)	$\mathbf{17.9}_{\pm \mathbf{0.15}}$	—

Table 21: Crop-window comparison. The default is $T_l \in [6,16]$, $T_g \in [32,48]$; the fixed-window rows use longer windows and larger frame budgets. The last row reproduces the live How2Sign entry of Table 2.

Visual backbone. The frozen frame backbone is varied while the temporal encoder, fusion module, decoder, and training schedule are fixed (Table 22). H2S BLEU ranges from 16.5 to 18.0 across five DINO backbones. DINOV3 ViT-B/16 reaches 17.9, while ViT-L/16 adds 0.1 BLEU at approximately 3.5 times the image-backbone parameters. These results support the B/16 default and show that the complete pipeline works across backbone sizes; geometry-specific effects are isolated by Table 6.

Frozen per-frame backbone	Params	H2S BLEU	BLEURT
DINOV2 ViT-S/14 (Oquab et al., 2024)	21M	16.5	49.4
DINOV2 ViT-L/14 (Oquab et al., 2024)	300M	17.4	50.5
DINOV3 ViT-S/16 (Siméoni et al., 2025)	22M	17.2	50.1
DINOV3 ViT-B/16 (default) (Siméoni et al., 2025)	86M	17.9	51.3
DINOV3 ViT-L/16 (Siméoni et al., 2025)	304M	18.0	51.4

Table 22: Visual backbone ablation with SIGNDEG live fine-tuned (same regime as the SIGNDEG live row of Table 2). The backbone is always frozen; only the per-frame CLS token is propagated to the SIGNDEGtemporal encoder. SSL pre-training is re-run per row; downstream translation training (Table 12) is identical across rows.

Translation decoder. Table 23 varies the decoder at the live fine-tuning operating point. Both upstreams use the source-to-target schedule of Table 12; decoder settings follow their releases.

SIGNDEG exceeds the SHUBERT comparison by 1.4–1.7 BLEU across these five decoders. ByT5-Base achieves the highest score in this comparison: ByT5-Large loses 0.4 BLEU for SIGNDEG, and mT5-Base/mBART-large trail by 0.8/1.1. This tests consistency across the evaluated decoders, rather than a decoder-independent ordering for all architectures.

Translation decoder	Params	SHUBERT BLEU	SIGNDEG BLEU
ByT5-Small (Xue et al., 2022)	300M	15.0	16.4
ByT5-Base (default) (Xue et al., 2022)	582M	16.2	17.9
ByT5-Large (Xue et al., 2022)	1.23B	15.9	17.5
mT5-Base (Xue et al., 2021)	580M	15.5	17.1
mBART-large (Liu et al., 2020)	610M	15.1	16.8

Table 23: Translation decoder ablation with both upstreams (SHUBERT and SIGNDEG) at live fine-tune. All rows use the same source-stage + target-stage recipe (Table 12); only the decoder backbone and its d_{model} projection change. The ByT5-Base default cells reproduce the corresponding SHUBERT (*16.2*) and SIGNDEG live (*17.9*) rows of Table 2.

Hand detector and hand geometry model. Table 24 varies the two hand-specific components of the pipeline under the matched-control protocol of Table 6. The upper block replaces the geometry model that supplies the GLP and GRT targets, with crops from YOLOv8n + ByteTrack. HaMeR (Pavlakos et al., 2024) has no MANO pose-query tokens, so it is compared with a WiLoR variant that also drops the latent target u_t and keeps θ_t and J_t. The lower block replaces the hand detector that produces the crops for both DINOV3 and the geometry model, with WiLoR targets.

I Ethics Statement

We use publicly released ASL corpora under their research terms and do not redistribute raw videos: YouTube-ASL and non-overlapping replacement clips from YouTube-SL-25 for source training; How2Sign, OpenASL, and FLEURS-ASL for translation; ASL Citizen, Sem-Lex, and WLASL2000 for recognition; and ASL-STEM-Wiki for fingerspelling. The frozen WiLoR model imports supervised hand-reconstruction knowledge. Aggregate benchmark scores do not establish equal performance across signers, skin tones, or signing styles. Face crops preserve linguistic non-manual signals and remain identifiable. Any deployment should address consent, privacy, and failure modes with Deaf-community participation (Desai et al., 2024).

Variant	H2S BLEU	ASLC R@1
Geometry model (detector: YOLOv8n + ByteTrack)		
None (SD only)	15.6	0.642
WiLoR, u_t, θ_t, J_t (default)	17.9	0.704
WiLoR, θ_t, J_t (no u_t)	16.8	0.670
HaMeR, θ_t, J_t	17.6	0.705
Hand detector (geometry: WiLoR)		
YOLOv8n + ByteTrack (default)	17.9	0.704
YOLOv8n only (no ByteTrack)	16.4	0.61
WiLoR detector + ByteTrack	17.5	0.694
MediaPipe Hands + ByteTrack	18.0	0.697

Table 24: Hand geometry model and hand detector substitutions under the matched-control protocol of Table 6 (984-hour source split, 450K updates, identical downstream schedules; H2S uses live SLT, ASLC full temporal fine-tuning). The default, SD-only, and no-ByteTrack rows reproduce Tables 6 and 7.

J Sample How2Sign Translations

Table 25 retains the supplied How2Sign translation examples alongside published SHUBERT hypotheses. Rows 1–6 show complete or shortened references; row 7 groups short fragments and is illustrative, not one aligned utterance. These selected examples illustrate translation differences and do not define a quantitative evaluation set.

K ASL Phonological Feature Classification Details

Like spoken languages, American Sign Language can be described through a system of phonological features. These features capture essential components of sign formation, including hand configuration, movement pattern, body-relative location, and contact. Following SHUBERT Appendix C, we use the ASL-Lex 2.0 feature set and list below the label values used for the 16 classification heads. A few feature values appear in only one of Sem-Lex or ASL Citizen; we keep the same feature vocabulary as SHUBERT to preserve benchmark comparability.

Handshape. v, 5, y, h, open_b, c, baby_o, flat_h, o, l, 1, a, open_8, w, curved_5, d, flatspread_5, i, f, s, p, flat_b, curved_4, flat_o, g, open_e, 4, closed_b, bent_1, 3, flat_horns, goody_goody, flat_m, bent_v, flat_1, r, 8, curved_v, open_h, curved_1, horns, flat_ily, flat_n, bent_l, stacked_5, ily, e, flat_v,

#	Source	Sentence
(1)	Reference	*And that's a great vital point technique for women's self defense.*
	SHUBERT	*This is a really great point for self defense.* (verbatim from Gueuwou et al. (2025b) Tab. 11)
	SIGNDEG	*And that's a great vital technique for women's self defense.*
(2)	Reference	*In this clip I'm going to show you how to tape your cables down.*
	SHUBERT	*In this clip I'm going to show you how to brand out the cable strings.*
	SIGNDEG	*In this clip I'm going to show you how to tape down your cables.*
(3)	Reference	*In this segment we're going to talk about how to load your still for distillation of lavender essential oil.*
	SHUBERT	*In this clip we're going to talk about how to take our stick for disinfectant oil.*
	SIGNDEG	*In this clip we're going to talk about how to load your still for the distillation of lavender oil.*
(4)	Reference	*You are dancing, and now you are going to need the veil and you are going to just grab the veil as far as possible.*
	SHUBERT	*Her dancing and now now she needs her feather to grab it with her foot as far as possible.*
	SIGNDEG	*You are dancing and now you need the veil, and you're going to grab the veil as far as possible.*
(5)	Reference	*But if you have to setup a new campfire, there's two ways to do it in a very low impact; one is with a mound fire ...*
	SHUBERT	*But if you have to set a new campfire, there are two ways to do a low impact one ...*
	SIGNDEG	*But if you have to set up a new campfire, there are two low-impact ways to do it; one is with a mound fire.*
(6)	Reference	*So, this is a very important part of the process.*
	SHUBERT	*This is a very important part of the process.*
	SIGNDEG	*So, this is a very important part of the process.*
(7)	Reference	*thank you / come on / now I've come this far ...* (representative short How2Sign segment)
	SHUBERT	*thank you / come on / how do you feel about it*
	SIGNDEG	*thank you / come on / now I have come this far*

Table 25: Selected qualitative How2Sign examples; row 7 contains grouped fragments. SHUBERT hypotheses are verbatim from Gueuwou et al. (2025b) Tables 11 and 13 (deduplicated and shortened).

curved_l, spread_open_e, curved_h, 7, closed_e, t, flat_4, open_f, k, and spread_e.

Nondominant Handshape. v, 5, y, none, open_b, Dominance Condition Violation, B, 1, a, open_8, C, s, h, o, flat_b, curved_5, p, c, S, closed_b, 4, flat_m, bent_v, flat_1, flat_h, baby_o, curved_v, i, f, bent_1, Symmetry Violation, flatspread_5, flat_o, curved_1, open_h, stacked_5, g, l, bent_l, 3, 8, spread_open_e, e, horns, w, r, Lax, curved_l, open_e, flat_4, O, curved_b, A, ily, flat_v, and flat_horns.

Minor Location. Neutral, Head Away, Body Away, Hand Away, Palm, Finger Tip, Forehead, Finger Front, Mouth, Chin, Other, Upper Arm, Torso Top, Forearm Back, Cheek Nose, Wrist Front, Palm Back, Finger Back, Finger Radial, Under Chin, Finger Ulnar, Wrist Back, Shoulder, Arm Away, Forearm Ulnar, Torso Mid, Heel, Clavicle, Eye, Forearm Front, Neck, Torso Bottom, Upper Lip, Head Top, Elbow Back, Hips, and Waist.

Second Minor Location. Neutral, Head Away, Torso Bottom, Finger Tip, Hand Away, none, Palm, Forearm Back, Finger Back, Body Away, Torso Top, Finger Front, Chin, Arm Away, Upper Arm, Finger Ulnar, Eye, Hips, Neck, Palm Back, Forearm Front, Finger Radial, Mouth, Heel, Torso Mid, Other, Waist, Cheek Nose, Forehead, Elbow Back, Under Chin, Clavicle, Shoulder, Forearm Ulnar, Head Top, Upper Lip, and Forearm Radial.

Sign Type. Symmetrical Or Alternating, One Handed, Dominance Violation, Asymmetrical Different Handshape, Asymmetrical Same Handshape, and Symmetry Violation.

Path Movement. Curved, Back And Forth, Straight, Circular, None, Z-shaped, Other, and X-shaped.

Flexion. Fully Open, Curved, Bent, Flat, none, Fully Closed, Stacked, and Crossed.

Selected Fingers. im, imrp, p, i, t, m, ip, imp, mr, imr, r, and mrp.

Major Location. Neutral, Head, Body, Hand, and Arm.

Spread Change. 1.0, 0.0, and none.

Thumb Contact. 1.0, 0.0, and none.

Spread. 1.0, 0.0, and none.

Thumb Position. Closed and Open.

Repeated Movement. 1.0 and 0.0.

Contact. 1.0 and 0.0.

Wrist Twist. 0.0 and 1.0.

Interpretation. These classifiers decompose performance into linguistically meaningful categories. Location and contact assess spatial cues; handshape and finger features assess manual articulation; path movement, repetition, and wrist twist assess temporal properties. They evaluate labels independent of WiLoR. Attribution to GLP and GRT is tested by the matched objectives and cross-signer retrieval rather than by comparing only to an external SSL system.

L Discussion

Why geometry enters an intermediate block. The temporal appearance pathway provides contextual frame features. GLP at block 4 supervises

those features with an articulated target while leaving two subsequent blocks for the sequence and relational objectives. The weighted downstream sum can use both intermediate and final states. The block-6 control in Table 6 tests whether supervising only the final representation loses some of this benefit. The design specifies an interface for geometry within temporal learning, rather than assuming that encoder depth automatically corresponds to a linguistic hierarchy.

Relationship to SHuBERT and SignDINO. SHuBERT learns discrete hidden-unit targets across signing streams, and the CVPR SignDINO (Gan et al., 2026) learns from frame-level articulator-focused views. SignDEG uses continuous temporal self-distillation and explicit hand-model targets. Published results contextualise task quality; shared evaluation heads do not make different upstreams identical experiments. The matched controls in Table 6 hold the appearance backbone and schedule fixed, directly testing geometry's contribution within SignDEG.

Interpreting the smaller live fine-tuning gain. Frozen and live SignDEG differ by 1.1 BLEU on How2Sign (16.8 to 17.9). Frozen features are useful without updating the temporal encoder; live fine-tuning improves task quality further. This within-model comparison isolates downstream adaptation and does not compare adaptation gains from different published ablation protocols.