Evaluation of Pose Estimation Systems for Sign Language Translation

Catherine O'Brien*, Gerard Sant*, Mathias Müller, Sarah Ebling
Department of Computational Linguistics
University of Zurich, Switzerland
{catherineelizabeth.obrien, gerard.santmuniesa}@uzh.ch, {mmueller, ebling}@cl.uzh.ch

Abstract

Many sign language translation (SLT) systems operate on pose sequences instead of raw video to reduce input dimensionality, improve portability, and partially anonymize signers. The choice of pose estimator is often treated as an implementation detail, with systems defaulting to widely available tools such as MediaPipe Holistic or OpenPose. We present a systematic comparison of pose estimators for pose-based SLT, covering widely used baselines (MediaPipe Holistic, OpenPose) and newer whole-body/high-capacity models (MMPose WholeBody, OpenPifPaf, AlphaPose, SDPose, Sapiens, SMPLest-X). We quantify downstream impact by training a controlled SLT pipeline on RWTH-PHOENIX-Weather 2014 where only the pose representation varies, evaluating with BLEU and BLEURT. To contextualize translation outcomes, we analyze temporal stability, missing hand keypoints, and robustness to occlusion using higher-resolution videos from the Signsuisse dataset. SDPose and Sapiens achieve the best translation performance (BLEU ~11.5), outperforming the common MediaPipe baseline (BLEU ~10). In occlusion cases, Sapiens is correct in all tested instances (15/15), while OpenPifPaf fails in nearly all (1/15) and also yields the weakest translation scores. Estimators that frequently leave out hand keypoints are associated with lower BLEU/BLEURT. We release code that can be used not only to reproduce our experiments, but also considerably lowers the barrier for other researchers to use alternative pose estimators.

Keywords: sign language, sign language processing, machine translation, pose estimation

1. Introduction

Sign language processing (SLP) is gaining ground within Natural Language Processing (NLP), yet it remains substantially underrepresented (Bragg et al., 2019; Yin et al., 2021; Müller et al., 2022). In spoken-language NLP, many core modeling decisions and preprocessing choices have been systematically studied and benchmarked. In contrast, for sign languages, even fundamental design decisions remain underexplored. For instance, while tokenizer selection in spoken-language translation is informed by extensive prior work, it remains unclear how sign language data should best be preprocessed for sign language translation (SLT).

SLT systems can accept input from a variety of modalities, including raw video (Zhou et al., 2023; Ye et al., 2024; Sant et al., 2025), glosses—semantic transcriptions of meaning—(Camgoz et al., 2018; Yin and Read, 2020; Zhou et al., 2021), form notations like SignWriting (Sutton, 1990; Jiang et al., 2023) or HamNoSys (Prillwitz and Zienert, 1990), which encode articulatory form, or poses (Zhang et al., 2024; Arib et al., 2025; Gan et al., 2025; Hwang et al., 2025). Poses are approximations of a signer's skeleton and movements as they sign.

Poses are a common choice for SLT because they are more viable than glosses or form notations (Müller et al., 2023b), and offer a lightweight interpretable alternative to raw video. They also provide some degree of anonymization when compared to signer videos (Battisti et al., 2024; Moryossef et al., 2025). Poses may also be more signer-independent for low-resource sign languages (Holmes et al., 2022). Despite these advantages, video-based models still outperform pose-based systems (Moryossef et al., 2021; Tarrés et al., 2023; Sant et al., 2025), suggesting that current pose representations may lose critical linguistic information. Improving pose estimation quality therefore represents a promising pathway toward narrowing this performance gap.

Although numerous pose estimation systems can be applied to sign language data, their impact on downstream translation performance remains largely unexplored. In practice, most translation pipelines rely on MediaPipe (Lugaresi et al., 2019) or OpenPose (Cao et al., 2019), primarily due to accessibility and ease of integration rather than demonstrated suitability for sign language.

To our knowledge, this work presents the first systematic evaluation of pose estimators for SLT. Beyond translation performance, we examine their behavior under pose estimation challenges intrinsic to signing, including occlusion, temporal instability, and missing hand detections. Our analysis provides practical guidance for selecting pose representations and highlights factors that influence downstream translation quality.

The main contributions of this paper are:

* These authors contributed equally.

Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages, pages 371–386
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

- A controlled comparison of eight prominent pose estimators on the RWTH-PHOENIX-Weather 2014 dataset,
- An analysis of pose estimation failure modes relevant to sign language data, including occlusion, temporal instability, and missing hands keypoints,
- An open-source framework extending the `pose-format` library (Moryossef et al., 2023b) with support for all evaluated pose estimators, simplifying integration and lowering barriers to their use in SLT pipelines.

Code and scripts to reproduce our experiments are available at https://github.com/ZurichNLP/multimodalhugs-pipelines.

2. Background

2.1. Pose Estimation

Pose estimation systems extract human skeletal keypoints from video, representing body, hand, and facial articulators as spatio-temporal trajectories. Modern pipelines typically rely on deep learning–based detectors such as OpenPose (Cao et al., 2019), MediaPipe Holistic (Lugaresi et al., 2019; Grishchenko and Bazarevsky, 2020), or SMPL-based models (Loper et al., 2015; Cai et al., 2023; Yin et al., 2026) to recover 2D or 3D joint configurations from RGB input. These representations provide a compact and interpretable encoding of human motion and are widely used in sign language processing (Wei et al., 2016; Li et al., 2020; Saunders et al., 2022; Joshi et al., 2025) and generation pipelines (Saunders et al., 2020; Xiao et al., 2020; Moryossef et al., 2023a; Arkushin et al., 2023).

In sign language contexts, accurate capture of fine-grained articulations—including handshape, orientation, movement, and facial expression—is essential, as linguistic meaning is conveyed through coordinated manual and non-manual signals (Liddell and Johnson, 1989; Johnson and Liddell, 2010; Pfau and Quer, 2010; Brentari, 2011; Sandler, 2012; Benitez-Quiroz et al., 2014; Yin et al., 2021). Errors in keypoint localization or tracking may therefore lead to the loss of critical articulatory information (Moryossef et al., 2021; Moryossef, 2024), degrading the quality of representations used by downstream tasks.

Despite these challenges, sign language processing pipelines commonly rely on general-purpose pose estimators that are not optimized for sign language data, as noted in (Coster et al., 2023; Jiang et al., 2025). Recent advances in whole-body and 3D pose estimation offer improved hand and body modeling (Jin et al., 2020a; Zhu et al., 2023). While pose-based representations are widely used in SLT, the implications of estimator choice for translation quality remain insufficiently understood.

2.2. Sign language translation

Sign language translation (SLT) seeks to generate spoken-language text from signed input (Müller et al., 2022; De Coster et al., 2024). SLT systems have employed a wide range of input representations, spanning continuous visual signals—such as raw video (Zhou et al., 2023; Ye et al., 2024) or learned visual features (Tarrés et al., 2023; Gueuwou et al., 2025), among others—as well as discrete representations including gloss annotations (Camgoz et al., 2018; Yin and Read, 2020; Zhou et al., 2021) and form notations (Jiang et al., 2023).

While raw video inputs often achieve the strongest translation performance among modalities (Sant et al., 2025), pose-based representations have been widely adopted in SLT pipelines (Zhang et al., 2024; Arib et al., 2025; Gan et al., 2025; Hwang et al., 2025), as they provide a lower-dimensional alternative that preserves articulatory structure. However, the effectiveness of pose-based SLT may be limited by the fidelity of the underlying pose representations, which varies across commonly used general-purpose estimators (Section 2.1).

Comparative studies of pose estimators for sign language tasks remain limited and report mixed findings. Several works report stronger performance of MediaPipe relative to OpenPose (Moryossef et al., 2021; Müller et al., 2022, 2023c), a trend further confirmed by Coster et al. (2023), who also found MediaPipe to outperform MMPose (Jin et al., 2020b). In contrast, Lazo-Quispe et al. (2022) reported improved results using an MMPose-based pipeline. The impact of newer whole-body and high-capacity estimators on SLT performance has yet to be systematically examined.

3. Pose Estimators

As shown in Table 1, we consider both pose estimators widely used in previous sign language processing research—such as MediaPipe (Lugaresi et al., 2019) and OpenPose (Cao et al., 2019)—as well as more recent systems that have not been extensively used but appear to be strong candidates. All evaluated methods are human pose estimators rather than models specialized for sign language. We selected both top-down and bottom-up approaches to pose estimation. We exclude common pose estimators that do not estimate hands, such as YOLO (Maji et al., 2022) or PoseFormer (Zheng et al., 2021).

estimator	# keypoints	scheme	2D/3D	speed (fps)	GPU only	confidence
MediaPipe Holistic (Lugaresi et al., 2019)	576	-	3D	0.89 / 3.15	-	✔
OpenPose (Cao et al., 2019)	137	-	2D	- / 4.40	✔	✔
MMPose Wholebody (Jin et al., 2020b)	133	COCO Wholebody	2D	0.89 / 3.81	-	✔
OpenPifPaf (Kreiss et al., 2021)	133	COCO Wholebody	2D	1.21 / 4.42	-	✔
SDPose (Liang et al., 2025)	133	COCO Wholebody	2D	0.07 / 0.84	-	✔
Sapiens (Khirodkar et al., 2025)	308	Sapiens-308	2D	0.04 / 3.29	-	✔
AlphaPose (Fang et al., 2023)	136	Halpe-FullBody	2D	- / 22.89	✔	✔
SMPLest-X (Yin et al., 2026)	137	SMPL-X	2D	- / 8.36	✔	-

Table 1: General characteristics of pose estimators considered in this study. scheme=whether estimated keypoints correspond to a standard layout, speed=frames per second on CPU/GPU measured on a single V100, confidence=whether the estimator outputs confidence values for each keypoint

Mediapipe Holistic (Lugaresi et al., 2019) This system employs a unified architecture to jointly estimate body pose, facial landmarks, and detailed hand keypoints by integrating region-specific subnetworks within a graph-based perception pipeline. This design produces dense 540-landmark representations and enables coordinated tracking of upper-body articulators from RGB input, making it widely used in real-time gesture and sign language processing. Its emphasis on low-latency, on-device inference has further contributed to its widespread adoption in SLT pipelines (Zhang et al., 2024; Gueuwou et al., 2025). In our experiments we reduce the full mesh of face keypoints to a smaller selection of contour points.

OpenPose (Cao et al., 2019) Building on Part Affinity Fields (PAFs) (Cao et al., 2017), which represents limb associations as 2D vector fields linking detected joints, this approach enables bottom-up multi-person pose estimation by jointly detecting keypoints and learning their pairwise associations. This formulation removes the dependency on person detection bounding boxes and improves robustness in crowded scenes. OpenPose further extended the pose estimation paradigm to full-body configurations including face and hands, which made it influential in multimodal and sign language research (Ko et al., 2019; Park and Sohn, 2020; Moryossef et al., 2020; Liang et al., 2023; Núñez-Marcos et al., 2023).

We run a pre-built OpenPose docker image with `--hands` and `--face` enabled, resulting in a total of 137 keypoints per frame. Occasionally OpenPose incorrectly predicts several people, in which case we select the first one.

MMPose Wholebody (Jin et al., 2020b) Extends conventional human pose estimation beyond standard 17-keypoint body configurations to dense full-body representations, typically comprising 133 keypoints (including body, hands, face, and feet) as defined in the COCO-WholeBody benchmark (Jin et al., 2020b). Architecturally, it follows a top-down paradigm: person instances are first detected, after which high-resolution keypoint heatmaps are predicted for each region.

We do estimation for MMPose Wholebody via their Unified Inference API. This API offers several checkpoint options and a variety of outputs. "Wholebody" refers to the checkpoint `rtmpose-m_simcc-coco-wholebody_pt-aic-coco_270e-256x192-cd5e845c_20230123`, which outputs the 133 keypoints defined in the COCO-Wholebody dataset.

OpenPifPaf (Kreiss et al., 2021) A bottom-up multi-person pose estimator builds on the Part Affinity Fields paradigm to jointly detect and associate keypoints. Its representation improves keypoint localization and grouping under partial occlusion and scale variation. OpenPifPaf has demonstrated strong robustness in crowded scenes (Andriluka et al., 2018) while remaining computationally efficient. We estimate poses for OpenPifPaf using their Predictor API with the pretrained checkpoint `shufflenetv2k30-wholebody`. We estimate poses for OpenPifPaf using their Predictor API with the pretrained checkpoint `shufflenetv2k30-wholebody`.

SDPose (Liang et al., 2025) Proposes a pose estimation framework that exploits pre-trained generative priors from Stable Diffusion (Rombach et al., 2022) to enhance both standard accuracy and robustness under domain shift. Evaluated against both in-domain and out-of-distribution benchmarks (Jin et al., 2020a; Ju et al., 2023) (e.g., human and stylized images), SDPose achieves competitive results with strong cross-domain generalization, highlighting the potential of diffusion-based priors in structured vision tasks such as pose estimation.

For SDPose, we use a pretrained model obtained from Hugging Face `teemosliang/SDPose-Wholebody`. This architecture first computes bounding boxes for each person using `YOLO-11x`. We have adapted their code such that we use only the person with the highest confidence, and discard any other persons detected.

Sapiens (Khirodkar et al., 2025) Represents a shift toward large-scale pretraining for human-centric vision tasks. Rather than optimizing narrowly for 2D keypoint detection, it leverages foundation-model-style pretraining on large-scale human motion data and supports fine-tuning for pose, segmentation, and depth estimation. Its dense 308-keypoint representation—including 243 facial landmarks—and high-resolution inference enable precise capture of fine-grained articulations relevant to sign language (See Section 2.1). This paradigm aligns with trends in NLP, where large-scale pretraining improves transferability to downstream multimodal tasks.

We estimate Sapiens poses based on the implementation of Gorordo (2024), which reproduces the official Sapiens pose model without requiring the full framework. We use the `sapiens_1b_goliath_best_goliath_AP_640` checkpoint from the official release; following the authors' recommendations, the 1B model is preferred as smaller variants yield lower accuracy.

AlphaPose (Fang et al., 2023) A top-down multi-person pose estimator that performs human detection followed by pose regression within each bounding box. It improves localization accuracy through Symmetric Integral Keypoint Regression, enabling continuous joint prediction beyond discrete heatmap representations. AlphaPose propose the Halpe-FullBody scheme, which supports whole-body estimation (up to 136 keypoints), extending COCO-WholeBody format by including additional head, neck, and hip joints. Its multi-stage pipeline enables efficient inference while preserving localization accuracy.

We use the authors' recommended `multi_domain_fast50_dcn_combined_256x192` checkpoint for Halpe-FullBody pose estimation.

SMPLest-X (Yin et al., 2026) A minimalist one-stage model for expressive human pose and shape estimation based on the SMPL-X scheme (Pavlakos et al., 2019) that jointly predicts body, hand, and face parameters using a transformer encoder–decoder architecture with task tokens, eliminating part-specific modules used in prior pipelines (Rong et al., 2021; Zhang et al., 2023). Designed to study scaling effects rather than architectural complexity, it is trained on large, diverse multi-dataset mixtures and shows strong cross-domain generalization. Despite its simplicity, the authors report state-of-the-art performance across multiple benchmarks (Pavlakos et al., 2019; Patel et al., 2021; Zhang et al., 2022; Lin et al., 2023; Fan et al., 2023) and highlight strong results for articulated hand pose estimation.

We use the `SMPLest-X-Huge` model recommended by the authors. Because this model does not provide per-keypoint confidence scores, we assume confidence 1 to retain all predicted keypoints during training (Section 4.1).

Support for `pose-format` We provide instructions and code[1] for running all of these lesser-used (in the context of SLT) estimators and also extend the `pose-format` library (Moryossef et al., 2023b) to support these new pose types. This considerably lowers the barriers for other researchers to use alternative estimators. Compatibility with pose-format means the predicted poses can more easily be stored, loaded, manipulated and visualized.

4. Methodology of Experiments

This study compares multiple pose estimators in the context of sign language translation. We evaluate each estimator within an identical translation pipeline to measure its impact on translation quality (Section 4.1). To contextualize these results, we also analyze estimator behavior with respect to temporal instability, occlusion, and missing hand detections—common challenges in pose estimation (Section 4.2).

4.1. Translation

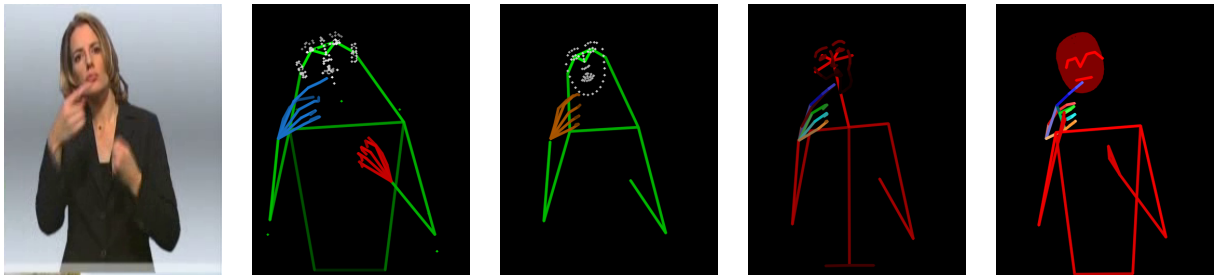
We train translation models with the recent MultiModalHugs toolkit (Sant et al., 2024). MultiModalHugs is a framework built on top of Hugging Face designed for multimodal AI models, which makes it ideal for a dataset of .pose files.

Dataset We use the RWTH-PHOENIX-Weather 2014 dataset (Forster et al., 2014; Camgoz et al., 2018)[2], which contains videos of German Sign language (DGS) aligned with German transcripts.

Preprocessing We process each video with each pose estimator to extract frame-level skeletal keypoints and store them in the binary .pose standard defined by the `pose-format` library. This enables consistent loading of pose sequences of any type within MultiModalHugs. The resulting representations preserve keypoint coordinates and confidence scores (if available) for each frame, while accommodating estimator-specific keypoint layouts (Table 1). Pose files are preprocessed as follows: leg keypoints are removed, and the remaining ones are spatially normalized per sequence to ensure consistent body and hand scale across signers. Missing values are zero-filled, and each frame is flattened into a feature vector.

[1] https://github.com/ZurichNLP/video-to-pose

[2] Referred to as *Phoenix* in the remainder of the text.



(a) Original frame (b) Sapiens (c) OpenPifPaf (d) OpenPose (e) MediaPipe

Figure 1: Example illustrating missing hand keypoints across pose estimators. (a) Original RGB frame used to extract poses, shown for reference. (b) Sapiens predicts a complete set of hand keypoints (no confidence $= 0$), although some finger articulation may still be inaccurate (e.g., the left index finger). (c–e) OpenPifPaf, OpenPose, and MediaPipe assign confidence $= 0$ to large portions of the hand keypoints (complete for c–d and nearly complete for e). During preprocessing, these keypoints are masked and zero-filled, resulting in the loss of hand information provided to the translation model.

We employ a shared encoder–decoder architecture, following the pose-based pipeline described in Sant et al. (2025). Pose sequences are represented as frame-level keypoint vectors and projected through a linear layer into the embedding space of a pretrained multilingual Transformer (`facebook/m2m100_418M`). This adapter aligns modality-specific representations with the language model's input format, ensuring that only the pose estimator varies across experiments. The input dimensionality of each model is determined by the number of keypoints per estimator, multiplied by their spatial dimensionality (2D or 3D).

Evaluation To evaluate translation quality, we report BLEU (Papineni et al., 2002) using SacreBLEU (Post, 2018)[3] and BLEURT (Sellam et al., 2020; Pu et al., 2021) using the `BLEURT-20` model. Other standard evaluation metrics such as COMET (Rei et al., 2020) are not applicable to our task, since they do not support sign languages.

4.2. Further Analyses

We conduct the following quantitative or qualitative analyses: an investigation of temporal jitter, robustness to occlusion and missing hand keypoints.

Datasets We use both the Phoenix and the Signsuisse dataset (Müller et al., 2023a). We add Signsuisse data because it includes higher-resolution videos than Phoenix and features deaf signers, whereas Phoenix has hearing interpreters. Besides, in order to accurately assess how estimators are handling occlusion, high-quality videos are required to reduce the influence of blur. We use the portion of the Signsuisse data that contains Swiss German Sign Language (*Deutschschweizerische Gebärdensprache*, or DSGS) videos.

[3]Signature: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.6.0`

Temporal jittering (pose stability) Beyond visual inspection, we quantify the temporal stability of each pose estimator using derivative-based smoothness metrics commonly adopted in video pose estimation. Following prior work (Zeng et al., 2022; Jin et al., 2023), we compute an acceleration-based jitter score (J_{acc}) as the mean magnitude of the second-order temporal difference of 2D keypoints, averaged over joints and time. To capture sharper temporal instabilities, we additionally compute a jerk-based score (J_{jerk}) using the third-order temporal difference, which penalizes rapid changes in acceleration and has been used as a smoothness indicator in motion and inertial pose estimation (Yi et al., 2021; Xia et al., 2022). We also report motion energy (E_v), defined as the mean joint velocity magnitude, to contextualize jitter values with respect to the overall amount of motion in the sequence. For 3D estimators, we compute these metrics using only the 2D image-plane coordinates.

Given the importance of manual and non-manual (e.g., facial) articulation in sign language, all metrics are computed on three subsets: (i) all keypoints (excluding legs), (ii) hands only, and (iii) face only. For each sequence, derivative magnitudes are averaged over joints and over time, yielding one scalar per sequence. Evaluation is conducted on 20 randomly selected videos from Signsuisse and 20 from Phoenix. Results are reported as median and interquartile range (IQR) across sequences (instead of mean and standard deviation), as these are more robust to the right-skewed nature of the observed distributions. Lower J_{acc} and J_{jerk} indicate smoother and more temporally stable predictions.

Occlusion Occlusion is a common source of error in pose estimation (Lino et al., 2025; Fan and Chowdhury, 2025), particularly in sign language, where one hand frequently overlaps the other or the face. To assess how estimators behave un-

Estimator	BLEU (↑)	BLEURT (↑)
MediaPipe	10.327 ± 0.269	0.351 ± 0.005
OpenPose	10.606 ± 0.251	0.353 ± 0.002
MMPose Wholebody	10.901 ± 0.299	0.361 ± 0.007
OpenPifPaf	9.365 ± 0.263	0.325 ± 0.007
SDPose	11.681 ± 0.415	0.372 ± 0.001
Sapiens	11.525 ± 0.222	0.372 ± 0.003
AlphaPose	11.251 ± 0.241	0.359 ± 0.007
SMPLest-X	9.709 ± 0.334	0.341 ± 0.009

Table 2: Translation scores on the Phoenix dataset. We report the mean and standard deviation across three training runs.

der these conditions, we analyzed 10 hand-picked videos from the Signsuisse dataset, comprising 15 individual signs that include occlusion.

We define *occlusion* as frames in which one hand or arm partially or fully obstructs the other hand or arm from the camera's viewpoint. We visually inspected the pose outputs for each estimator and evaluated whether both hands were consistently detected and whether their locations, handshapes, movements, and palm orientations remained plausible relative to the source video. A prediction was considered acceptable when both hands were present for most of the sequence and their articulatory configuration matched the observed signing. Given the limited sample size, this occlusion analysis is intended to contextualize the main translation results rather than provide a comprehensive robustness benchmark.

Missing hand keypoints During translation preprocessing, keypoints with confidence $c = 0$ are treated as invalid and zero-filled before frame-level flattening; such keypoints are therefore absent from the model input. Estimators that frequently produce $c = 0$ hand keypoints thus provide less hand information to the translation model. Example cases are shown in Figure 1. We therefore quantify missing hand keypoints directly from the `.pose` files to capture how often substantial portions of hand information are unavailable to the translation model.

For each frame and each hand, we compute the proportion of hand keypoints with $c = 0$. We define a hand as *missing* in a given frame when at least 50% of its keypoints have $c = 0$. Signing frames are defined as those where the wrist is vertically above the elbow.

5. Results & Discussion

5.1. Translation Results

Table 2 shows the evaluation results of our sign language translation (SLT) experiments. Every result is an average across three training runs. Overall, the translation quality in our experiments is much lower than state-of-the art systems, especially ones based on videos instead of poses, but we are interested specifically in the *relative* performance of pose estimators compared to each other.

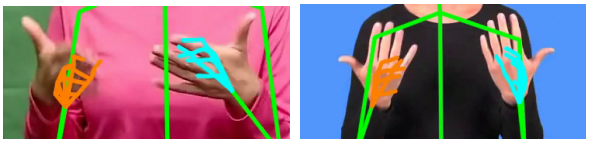


Figure 2: Examples of erroneous SMPLest-X hand pose estimates overlaid on original frames, cropped to highlight the hands. In both examples (left: How2Sign (Duarte et al., 2021); right: Signsuisse), the predicted hand keypoints exhibit incorrect orientation and articulation that do not match the observed hand configurations.

Sapiens and SDPose achieved the highest BLEU and BLEURT scores, of roughly 11.5 and 0.37, suggesting that they are stronger candidates for SLT applications compared to other estimators. In total, four different estimators (Sapiens, SDPose, MMPose Wholebody and AlphaPose) lead to better results than Mediapipe, the de-facto standard pose estimation system used in SLT.

This provides empirical evidence that **alternatives to Mediapipe may be better suited for SLT**, and **future work should therefore consider a broader range of pose estimation systems.**

5.2. Further Analyses

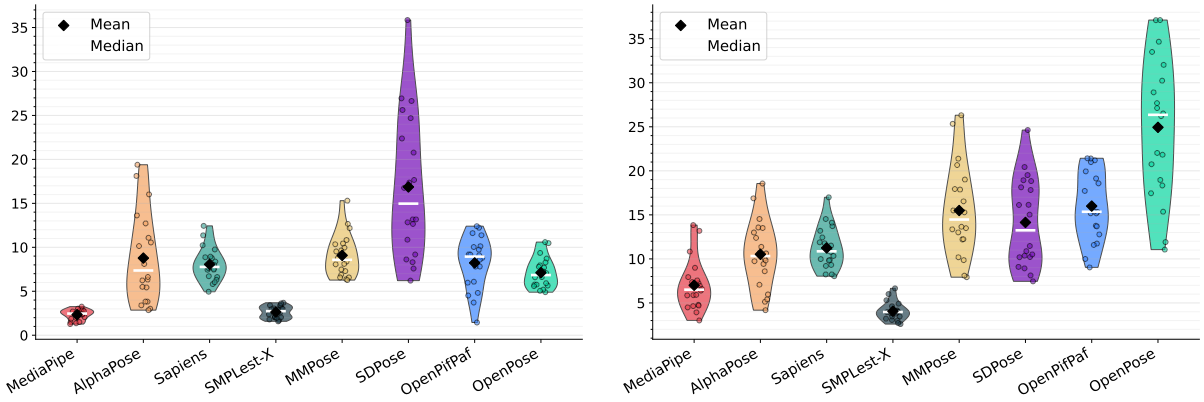
Temporal jittering (pose stability) Table 3 summarizes the temporal stability of pose estimators on Signsuisse and Phoenix for all keypoints (excluding legs); full results by anatomical region (hands, face) are in Appendix B. Overall, the quantitative trends align with visual inspection: methods that appear temporally unstable exhibit higher J_{acc} and J_{jerk}, particularly in the hand subset. While both metrics exhibit similar trends, J_{jerk} provides greater separation between estimators (Figure 3) and is thus used to illustrate differences.

On Signsuisse, MediaPipe achieves the lowest median jerk jitter for all keypoints ($J_{\text{jerk}} = 2.46$), followed closely by SMPLest-X (2.74). However, the violin plots reveal that their distributions overlap substantially (Figure 3a), indicating that the difference may not be reliable given the sample size of 20 sequences. SDPose exhibits the highest jerk values (14.97), with a wide interquartile range reflecting high variance across sequences, despite yielding the most stable face keypoints (see Appendix Table 8 for full results).

When focusing on hands (Appendix Table 7), SMPLest-X provides the lowest jitter ($J_{\text{jerk}} = 5.85$); however, this temporal smoothness coincides with visibly rigid hand configurations (Figure 2). For face

Pose Estimator	Signsuisse			Phoenix		
	E_v	J_{acc} (↓)	J_{jerk} (↓)	E_v	J_{acc} (↓)	J_{jerk} (↓)
MediaPipe	**1.16**	**1.44** (1.13–1.60)	**2.46** (2.01–2.81)	**3.07**	3.69 (3.01–4.73)	6.51 (5.10–8.38)
OpenPose	3.30	3.97 (3.56–4.74)	6.84 (5.90–8.18)	9.32	14.74 (11.38–16.64)	26.36 (19.57–29.55)
MMPose Wholebody	3.83	4.95 (4.26–5.90)	8.58 (7.32–10.35)	6.84	8.75 (7.00–11.17)	14.47 (11.66–18.59)
OpenPifPaf	2.76	4.84 (3.43–5.54)	8.94 (6.35–10.20)	5.22	8.60 (7.28–10.31)	15.35 (13.12–18.63)
SDPose	6.21	8.45 (6.19–11.76)	14.97 (10.70–21.79)	6.83	7.99 (6.61–10.45)	13.24 (10.80–17.10)
Sapiens	2.97	4.42 (3.72–5.05)	7.77 (6.62–9.16)	4.55	6.33 (5.44–7.43)	10.86 (9.62–12.88)
AlphaPose	4.09	4.56 (3.08–6.70)	7.36 (5.15–11.50)	5.81	6.21 (4.90–7.69)	10.32 (8.35–12.47)
SMPLest-X	2.13	1.74 (1.39–2.00)	2.74 (2.16–3.14)	3.08	**2.68** (2.23–3.30)	**3.97** (3.19–4.85)

Table 3: Temporal stability metrics on Signsuisse and Phoenix datasets, for all keypoints excluding legs. E_v: motion energy (median joint velocity). J_{acc} and J_{jerk}: acceleration- and jerk-based jitter; lower is smoother. Values are median across 20 sequences, with interquartile range (Q1–Q3) in parentheses. All values scaled by 100. Boldface marks the lowest value when distributions are clearly separated (see Figure 3). Full results by anatomical region in Appendix B.



(a) Signsuisse

(b) Phoenix

Figure 3: Distribution of per-sequence jerk jitter (J_{jerk}) across pose estimators for all keypoints (excluding legs). Each point represents one video sequence ($n = 20$). Black diamonds indicate the mean; white bars indicate the median. The violin shape shows the kernel density estimate. Overlapping distributions (e.g., MediaPipe vs. SMPLest-X on Signsuisse) indicate that differences between estimators may not be reliable at this sample size.

keypoints, SDPose achieves the lowest jitter (0.38), though MediaPipe (0.44) and OpenPifPaf (0.43) are nearly indistinguishable.

On Phoenix, instability increases for all methods, consistent with the lower spatial resolution of the videos (approximately 5.3× fewer pixels than Signsuisse). SMPLest-X achieves the lowest median jerk jitter for all keypoints (3.97) and hands (7.40), with distributions clearly separated from other estimators (Figure 3b). MediaPipe is second-lowest overall (6.51). OpenPose exhibits the highest jitter across all regions, with a median J_{jerk} of 26.36 for all keypoints and 67.74 for hands. For face keypoints, SDPose achieves the lowest jitter (1.17), while Sapiens is notably high (10.29), consistent with its dense 243-landmark face mesh amplifying localization noise at low resolution.

Across both datasets, the violin plots (Figure 3) provide important context beyond summary statistics: estimators with similar median values may have heavily overlapping distributions, tempering conclusions about which is "best." When interpreted jointly with motion energy, the results suggest that SMPLest-X favors temporal smoothness at the cost of limited articulation, while MediaPipe—and to a lesser extent Sapiens—offer a trade-off between stability and expressive motion.

Occlusion The results of the occlusion analysis are available in Table 4. Sapiens produced accurate estimations for 15/15 instances of occlusion surveyed. Notably, there are several circumstances where MediaPipe is missing one or both hands for a given sign, but Sapiens produces the pose correctly. One example of this trend is shown in Figure 4. OpenPifPaf, on the other hand, was missing some or all of one or both hands in all but one of the 15 instances of occlusion surveyed. This propensity

estimator	correctness (%)
Mediapipe	73.33
OpenPose	66.66
MMPose Wholebody	33.33
OpenPifPaf	6.66
SDPose	46.66
Sapiens	100.00
AlphaPose-136	40.00
SMPLest-X	0.00

Table 4: Percentage of the time that the estimators correctly estimated a pose featuring occlusion in the 15 Signsuisse examples surveyed.

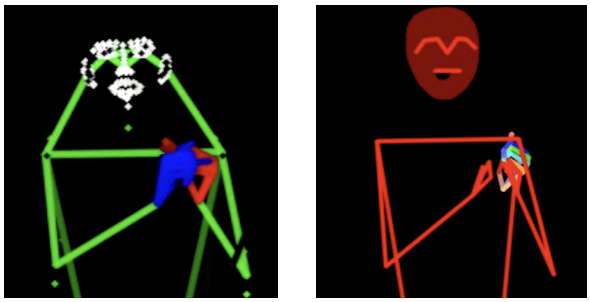


Figure 4: Pose estimation visualizations with hand occlusion for the DSGS sign SPORT. Left: Sapiens correctly predicts both hands despite partial occlusion. Right: MediaPipe fails to detect the right hand, resulting in missing keypoints. Images are cropped to highlight the hands.

for missing hands may contribute to the low BLEU score of OpenPifPaf. SMPLest-X, in turn, did provide hands in all frames, but did not produce any accurate representations of the pose.

MMPose Wholebody and SDPose both estimated a hand position for all frames in all of the 15 examples surveyed, but were more likely than other estimators to estimate an impossible hand position, for example bending the fingers backwards. MMPoseWholebody created an acceptable estimation for 5/15 instances of occlusion, while SDPose produced an acceptable estimation in 7/15 examples.

Missing hand keypoints We report the percentage of signing frames with a missing hand for each estimator (Table 5), computed on 20 randomly selected videos from the Phoenix training split. A full breakdown across varying missing thresholds (10–100%) is provided in Appendix A. Only OpenPifPaf, MediaPipe, and OpenPose assign $c = 0$ to hand keypoints; all other estimators never produce $c = 0$, resulting in 0% missing across all thresholds.

OpenPifPaf shows severe degradation: in roughly 68% of signing frames the left hand is missing and in 59% the right hand is missing. In over 40% of signing frames, both hands are simultaneously absent. MediaPipe exhibits a milder but consistent pattern, with around 20–24% of signing frames missing a hand. OpenPose is affected less frequently, with 3–8% of frames affected. Notably, both MediaPipe and OpenPose exhibit binary behavior: when a hand is lost, all of its keypoints are set to $c = 0$ at once rather than partially, as shown by the constant values across thresholds in the full breakdown (Table 6 in Appendix A).

estimator	left	right	both
MediaPipe	20.22	23.61	8.84
OpenPose	8.43	3.88	0.22
OpenPifPaf	67.65	59.19	40.63

Table 5: Percentage of signing frames with a missing hand on 20 randomly selected Phoenix training videos. A hand is missing when $\geq 50\%$ of its keypoints have $c = 0$. "both": both hands missing simultaneously. Only estimators producing $c = 0$ hand keypoints are shown; all others (AlphaPose, MMPose Wholebody, Sapiens, SDPose, SMPLest-X) score 0%. Full threshold sweep in Appendix A.

These three estimators also obtain some of the weakest translation scores (Table 2), suggesting that zero-confidence hand keypoints—explicitly masked during preprocessing—directly impair downstream translation. This pattern could also reveal a limited capability of current translation models to recover missing hand information from temporal context alone.

6. Conclusions

We presented a controlled comparison of pose estimators for pose-based SLT, motivated by the fact that most prior SLT pipelines default to MediaPipe as a convenient choice. Our experiments show that this default is not necessarily optimal: several estimators outperform MediaPipe on Phoenix, including SDPose, Sapiens, AlphaPose, and MMPose Wholebody. Among them, SDPose and Sapiens achieve the highest BLEU/BLEURT scores.

Further analyses on signing data Estimators that frequently remove substantial hand information (notably OpenPifPaf, and to a lesser extent MediaPipe and OpenPose) are associated with weaker translation results. The temporal stability analysis further shows that higher translation performance does not necessarily coincide with smoother trajectories: MediaPipe and SMPLest-X yield the lowest jitter scores, whereas SDPose exhibits comparatively high hand jerk on Signsuisse, and SMPLest-X attains very low jitter at the cost of visibly implausible or rigid hand articulations. Taken together, these findings underscore that estimator choice can materially affect both translation performance and pose quality characteristics relevant to signing.

Trade-off between translation quality and computational cost Sapiens provides strong translation performance and the most robust behavior under occlusion (15/15 correct in our manual analysis), but it is also among the most resource-intensive and slowest options. SDPose reaches comparable translation quality results with similarly high runtime requirements. Conversely, faster or more lightweight estimators are not guaranteed to be suitable for SLT: OpenPifPaf, for instance, performs poorly in translation and exhibits severe hand keypoint dropout. AlphaPose, which has the third-best BLEU score and the fastest compute speed, may be a better option when time or computing resources are limited.

Overall, our results provide empirical evidence that pose-based SLT should not treat pose estimation as a fixed implementation detail. Future research should therefore consider pose estimators beyond MediaPipe when building pose-based translation systems, and evaluate estimator choice jointly with practical constraints such as runtime and hardware requirements as well as robustness properties such as occlusion handling, missing-hand behavior, and temporal stability.

7. Limitations and Future Work

Limitations of the Phoenix dataset It should be noted that the signing in the Phoenix dataset is done live by hearing interpreters. Accordingly, this dataset has notable flaws. Due to the time pressure of the live setting, the interpreters may omit some information. Furthermore, the signing is an interpretation of German spoken language, and not natural signing, and thus may be influenced by German grammar. Lastly, the subject domain of Phoenix is fairly limited and mostly pertains to weather reports. Future investigations should expand the analysis of pose estimators to new datasets, particularly those that contain natural signing from deaf L1 signers with high-quality translations that are not produced under time pressure.

Translation experiments on more suitable, newer, or larger sign language datasets would provide a more reliable assessment of how pose estimator choice generalizes beyond Phoenix.

Fast movement Anecdotally, a visual analysis suggests that estimators such as Sapiens may be vulnerable to errors on videos containing fast movements (which is normal for signing). Future research could therefore investigate the estimators' robustness to fast movements.

Variable model capacity All experiments use a shared translation architecture with a linear projection layer that maps pose features into the language model embedding space (see Section 4.1). Because pose estimators produce different numbers of keypoints (Table 1), this projection layer contains a different number of parameters for each estimator. While this design allows each representation to be fully utilized, it introduces variation in model capacity that may influence translation performance. Isolating the impact of adapter dimensionality is left for future work.

Acknowledgements

CO is supported by a Fulbright Scholarship. GS and MM received funding from the SIGMA project (grant no. G-95017-01-07), supported by the Digital Society Initiative (DSI) at the University of Zurich.

8. Bibliographical References

Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. 2018. Posetrack: A benchmark for human pose estimation and tracking. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5167–5176.

Safaeid Hossain Arib, Rabeya Akter, Sejuti Rahman, and Shafin Rahman. 2025. Signformer-gcn: Continuous sign language translation using spatio-temporal graph convolutional networks. *PLOS ONE*, 20(2):1–19.

Rotem Shalev Arkushin, Amit Moryossef, and Ohad Fried. 2023. Ham2pose: Animating sign language notation into pose sequences. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21046–21056.

Alessia Battisti, Emma van den Bold, Anne Göhring, Franz Holzknecht, and Sarah Ebling. 2024. Person identification from pose estimates in sign language. *11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*.

C. Fabian Benitez-Quiroz, Kadir Gökgöz, Ronnie B. Wilbur, and Aleix M. Martinez. 2014. Discriminant features and temporal structure of nonmanuals in american sign language. *PLOS ONE*, 9(2):1–17.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, pages 16–31,

New York, NY, USA. Association for Computing Machinery.

Diane Brentari. 2011. Sign language phonology. In John Goldsmith, Jason Riggle, and Alan C. L. Yu, editors, *The Handbook of Phonological Theory*, 2nd edition, chapter 21, pages 691–721. John Wiley & Sons.

Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. 2023. Smpler-x: Scaling up expressive human pose and shape estimation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310.

Mathieu De Coster, Ellen Rushe, Ruth Holmes, Anthony Ventresque, and Joni Dambre. 2023. Towards the extraction of robust sign embeddings for low resource sign language recognition.

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Steven Verstockt. 2024. Machine translation from signed to spoken languages: State of the art and challenges. *Universal Access in the Information Society*, 23:1305–1331.

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 (CVPR)*, pages 2734–2743.

Chengyu Fan and Tahiya Chowdhury. 2025. When pose estimation fails: Measuring occlusion for reliable multimodal interaction. In *Companion Proceedings of the 27th International Conference on Multimodal Interaction*, pages 58–64.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. 2023. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 12943–12954. IEEE.

Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. 2023. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7157–7173.

Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1911–1916, Reykjavik, Iceland. European Language Resources Association (ELRA).

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Lei Xie, and Sanglu Lu. 2025. Vision-based sign language translation via a skeleton-aware neural network. *J. Comput. Sci. Technol.*, 40(2):378–396.

Ibai Gorordo. 2024. Sapiens pytorch inference. `https://github.com/ibaiGorordo/Sapiens-Pytorch-Inference`. GitHub repository.

Ivan Grishchenko and Valentin Bazarevsky. 2020. MediaPipe Holistic - simultaneous face, hand and pose prediction, on device. Google AI Blog.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, Karen Livescu, and Alexander H. Liu. 2025. SHuBERT: Self-supervised sign language representation learning via multi-stream cluster prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28792–28810, Vienna, Austria. Association for Computational Linguistics.

Ruth Holmes, Ellen Rushe, Frank Fowley, and Anthony Ventresque. 2022. Improving signer independent sign language recognition for low resource languages. In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 45–52, Marseille, France. European Language Resources Association.

Eui Jun Hwang, Sukmin Cho, Huije Lee, Youngwoo Yoon, and Jong C. Park. 2025. A spatio-temporal representation learning as an alternative to traditional glosses in sign language translation and production. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3352–3362.

Zifan Jiang, Colin Leong, Amit Moryossef, Oliver Cory, Maksym Ivashechkin, Neha Tarigopula, Biao Zhang, Anne Göhring, Annette Rios, Rico Sennrich, and Sarah Ebling. 2025. Meaningful pose-based sign language evaluation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 64–80, Suzhou, China. Association for Computational Linguistics.

Zifan Jiang, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2023. Machine translation between spoken languages and signed languages represented in SignWriting. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1706–1724, Dubrovnik, Croatia. Association for Computational Linguistics.

Kyung-Min Jin, Byoung-Sung Lim, Gun-Hee Lee, Tae-Kyung Kang, and Seong-Whan Lee. 2023. Kinematic-aware hierarchical attention network for human pose estimation in videos. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5714–5723.

Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. 2020a. Whole-body human pose estimation in the wild. In *Computer Vision – ECCV 2020*, pages 196–214, Cham. Springer International Publishing.

Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. 2020b. Whole-body human pose estimation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Robert E. Johnson and Scott K. Liddell. 2010. Toward a phonetic representation of signs: Sequentiality and contrast. *Sign Language Studies*, 11(2):241–274.

Abhinav Joshi, Vaibhav Sharma, Sanjeet Singh, and Ashutosh Modi. 2025. PoseStitch-SLT: Linguistically inspired pose-stitching for end-to-end sign language translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13834–13853, Suzhou, China. Association for Computational Linguistics.

Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. 2023. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 618–629.

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. 2025. Sapiens: Foundation for human vision models. In *Computer Vision – ECCV 2024*, pages 206–228, Cham. Springer Nature Switzerland.

Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9(13).

Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. 2021. Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):13498–13511.

Cristian Lazo-Quispe, Joe Huamani-Malca, Manuel Huamán-Ramos, Pablo Rivas, and Tomas Cerny. 2022. Impact of pose estimation models for landmark-based sign language recognition. In *LXAI Workshop Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS 2022)*.

Dongxu Li, Cristian Rodriguez Opazo, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1448–1458.

Shuang Liang, Jing He, Chuanmeizhi Wang, Lejun Liao, Guo Zhang, Yingcong Chen, and Yuan Yuan. 2025. Sdpose: Exploiting diffusion priors for out-of-domain and robust pose estimation.

Zeyu Liang, Huailing Li, and Jianping Chai. 2023. Sign language translation: A survey of approaches and techniques. *Electronics*, 12(12).

Scott K Liddell and Robert E Johnson. 1989. American sign language: The phonological base. *Sign language studies*, 64(1):195–277.

Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. 2023. One-stage 3d whole-body mesh recovery with component aware transformer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21159–21168.

Filipa Lino, Carlos Santiago, and Manuel Marques. 2025. Benchmarking 3d human pose estimation models under occlusions. *arXiv preprint arXiv:2504.10350*.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6).

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. *CoRR*, abs/1906.08172.

Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. 2022. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2637–2646.

Amit Moryossef. 2024. Optimizing hand region detection in mediapipe holistic full-body pose estimation to improve accuracy and avoid downstream errors.

Amit Moryossef, Mathias Müller, Anne Göhring, Zifan Jiang, Yoav Goldberg, and Sarah Ebling. 2023a. An open-source gloss-based baseline for spoken to signed language translation. In *Proceedings of the Second International Workshop on Automatic Translation for Signed and Spoken Languages*, pages 22–33, Tampere, Finland. European Association for Machine Translation.

Amit Moryossef, Mathias Müller, and Rebecka Fahrni. 2023b. pose-format: Library for viewing, augmenting, and handling .pose files.

Amit Moryossef, Gerard Sant, and Zifan Jiang. 2025. Pose-based sign language appearance transfer. In *Proceedings of the Third International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 1–6, Geneva, Switzerland. European Association for Machine Translation.

Amit Moryossef, Ioannis Tsochantaridis, Roee Aharoni, Sarah Ebling, and Srini Narayanan. 2020. Real-time sign language detection using human pose estimation. In *Computer Vision - ECCV 2020 Workshops - Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, volume 12536 of *Lecture Notes in Computer Science*, pages 237–248. Springer.

Amit Moryossef, Ioannis Tsochantaridis, Joe Dinn, Necati Cihan Camgoz, Richard Bowden, Tao Jiang, Annette Rios, Mathias Muller, and Sarah Ebling. 2021. Evaluating the immediate applicability of pose estimation for sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3434–3440.

Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-Bonet, Anne Göhring, Roman Grundkiewicz, Mert Inan, Zifan Jiang, Oscar Koller, Amit Moryossef, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi, and Davy Van Landuyt. 2023a. Findings of the second WMT shared task on sign language translation (WMT-SLT23). In *Proceedings of the Eighth Conference on Machine Translation*, pages 68–94, Singapore. Association for Computational Linguistics.

Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-Bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, and Katja Tissi. 2022. Findings of the first WMT shared task on sign language translation (WMT-SLT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 744–772, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023b. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Amit Moryossef, Sarah Ebling, and Thomas Hanke. 2023c. EASIER project deliverable 4.3: Final translation systems V2.

Adrián Núñez-Marcos, Olatz Perez de Viñaspre, and Gorka Labaka. 2023. A survey on sign language machine translation. *Expert Systems with Applications*, 213:118993.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Chan-Il Park and Chae-Bong Sohn. 2020. Data augmentation for human keypoint estimation deep learning based sign language translation. *Electronics*, 9(8).

Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. 2021. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13463–13473.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10967–10977.

Roland Pfau and Josep Quer. 2010. Nonmanuals: Their grammatical and prosodic roles. In Diane Brentari, editor, *Sign Languages*, Cambridge Language Surveys, pages 381–402. Cambridge University Press, Cambridge.

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. Association for Computational Linguistics.

Siegmund Prillwitz and Heiko Zienert. 1990. Hamburg notation system for sign language: Development of a sign writing with computer application. In *Current Trends in European Sign Language Research: Proceedings of the 3rd European Congress on Sign Language Research*, pages 355–379.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685.

Yu Rong, Takaaki Shiratori, and Hanbyul Joo. 2021. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1749–1759.

Wendy Sandler. 2012. The phonological organization of sign languages. *Language and Linguistics Compass*, 6(3):162–182.

Gerard Sant, Zifan Jiang, Carlos Escolano, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. Multimodalhugs: Enabling sign language processing in hugging face. Manuscript submitted for publication.

Gerard Sant, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2025. Modality matters: Training and tokenization effects in sign-to-text translation. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, IVA Adjunct '25, New York, NY, USA. Association for Computing Machinery.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2020. Progressive transformers for end-to-end sign language production. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, volume 12356 of *Lecture Notes in Computer Science*, pages 687–705. Springer.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2022. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5131–5141.

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, Online. Association for Computational Linguistics.

Valerie Sutton. 1990. *Lessons in SignWriting*. SignWriting.

Laia Tarrés, Gerard I. Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró-i Nieto. 2023. Sign language translation from instructional videos. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5625–5635.

Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *2016 IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR), pages 4724–4732.

Di Xia, Yeqing Zhu, and Heng Zhang. 2022. Faster deep inertial pose estimation with six inertial sensors. *Sensors*, 22(19).

Qinkun Xiao, Minying Qin, and Yuting Yin. 2020. Skeleton-based chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural Networks*, 125:41–55.

Jinhui Ye, Xing Wang, Wenxiang Jiao, Junwei Liang, and Hui Xiong. 2024. Improving gloss-free sign language translation by reducing representation density. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Xinyu Yi, Yuxiao Zhou, and Feng Xu. 2021. Transpose: real-time 3d human translation and pose estimation with six inertial sensors. *ACM Trans. Graph.*, 40(4).

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

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, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wanqi Yin, Zhongang Cai, Ruisi Wang, Ailing Zeng, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, Hui En Pang, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Atsushi Yamashita, Lei Yang, and Ziwei Liu. 2026. Smplest-x: Ultimate scaling for expressive human pose and shape estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):1778–1794.

Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. 2022. Smoothnet: A plug-and-play network for refining human poses in videos. In *Computer Vision – ECCV 2022*, pages 625–642, Cham. Springer Nature Switzerland.

Biao Zhang, Garrett Tanzer, and Orhan Firat. 2024. Scaling sign language translation. *Advances in neural information processing systems*, 37:114018–114047.

Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. 2023. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12287–12303.

Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. 2022. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI*, volume 13666 of *Lecture Notes in Computer Science*, pages 180–200. Springer.

Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 2021. 3d human pose estimation with spatial and temporal transformers. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 20814–20824. IEEE.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325.

Yue Zhu, Nermin Samet, and David Picard. 2023. H3wb: Human3.6m 3d wholebody dataset and benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20166–20177.

Estimator	Hand	% hand keypoints missing ($\geq x$)									
		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
OpenPifPaf	Left	85.16	80.86	76.51	72.10	67.65	63.29	59.31	54.59	51.81	0.00
	Right	82.59	75.01	69.90	63.68	59.19	55.66	51.96	49.01	46.19	0.00
	Both	66.90	57.98	52.14	45.91	40.63	36.43	32.32	28.91	26.05	0.00
MediaPipe	Left	20.22	20.22	20.22	20.22	20.22	20.22	20.22	20.22	20.22	0.00
	Right	23.61	23.61	23.61	23.61	23.61	23.61	23.61	23.61	23.61	0.00
	Both	8.84	8.84	8.84	8.84	8.84	8.84	8.84	8.84	8.84	0.00
OpenPose	Left	8.43	8.43	8.43	8.43	8.43	8.43	8.43	8.43	8.43	0.00
	Right	3.88	3.88	3.88	3.88	3.88	3.88	3.88	3.88	3.88	0.00
	Both	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.00

Table 6: Percentage of signing frames in which at least $x\%$ of a hand's keypoints have confidence $c = 0$, measured on 20 randomly selected videos from the Phoenix training split. Signing frames are those where the wrist is vertically above the elbow. "Both": both hands simultaneously exceed the threshold. Only estimators producing $c = 0$ hand keypoints are shown; all others score 0%. The 50% column corresponds to the missing hand definition in Table 5.

A. Missing Hand Keypoints

Table 6 reports the percentage of signing frames with missing hand keypoints across varying thresholds. The 50% column corresponds to the definition of a missing hand used in the main text (Table 5). MediaPipe and OpenPose exhibit binary behavior: when a hand is lost, all its keypoints are set to $c = 0$ simultaneously, resulting in constant values across thresholds 10–90%.

B. Temporal Stability by Region

Tables 7 and 8 report temporal stability metrics for hand and face keypoints on Signsuisse and Phoenix. The all-keypoints region is reported in the main text (Table 3). Corresponding violin plots are shown in Figures 5 and 6.

Hands Hand keypoints exhibit substantially higher jitter than full-body or face keypoints across all estimators, reflecting the difficulty of localizing fine-grained finger articulations.

On Signsuisse, SMPLest-X achieves the lowest median hand jerk jitter ($J_{\text{jerk}} = 5.85$), clearly separated from the next-lowest estimator (MMPose Wholebody, 10.48). However, as noted in the main text, this smoothness coincides with rigid, often implausible hand configurations (Figure 2). SDPose exhibits the highest hand jitter (29.98) with very wide interquartile range, indicating high variance across sequences. MediaPipe (25.29) and OpenPifPaf (22.61) also show high hand jitter on Signsuisse, though for different reasons: MediaPipe's hand jitter is inflated by its frequent missing-hand detections (Table 5), while OpenPifPaf's is driven by noisy localization even when hands are detected.

On Phoenix, the same pattern holds with larger values: SMPLest-X remains lowest (7.40), while

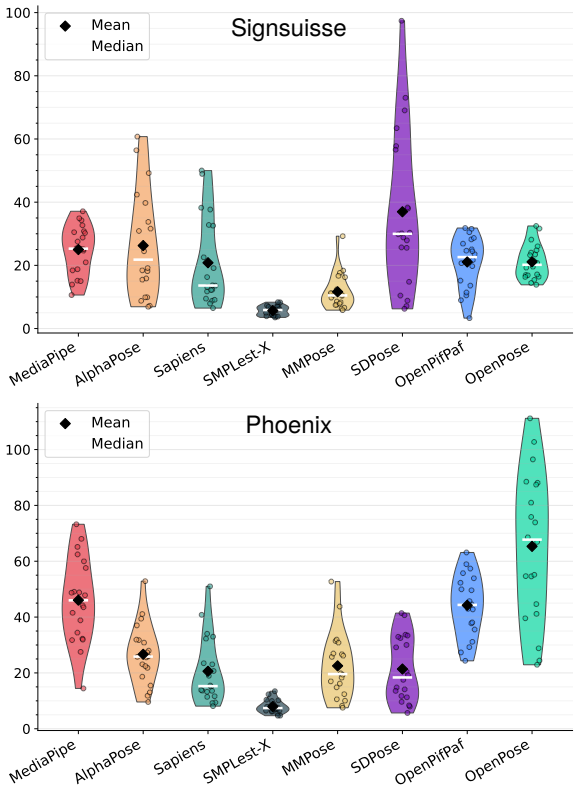


Figure 5: Distribution of per-sequence jerk jitter (J_{jerk}) for hand keypoints on Signsuisse (top) and Phoenix (bottom). Format as in Figure 3.

OpenPose (67.74) and MediaPipe (45.99) are the most unstable. The increase relative to Signsuisse is consistent with Phoenix's lower resolution making hand keypoint localization more challenging.

Face Face keypoints show the lowest jitter overall, as facial landmarks undergo less motion than hands or body during signing. Differences between estimators are smaller but still informative.

Pose Estimator	Signsuisse			Phoenix		
	E_v	J_{acc} (↓)	J_{jerk} (↓)	E_v	J_{acc} (↓)	J_{jerk} (↓)
MediaPipe	10.49	14.50 (10.73–16.23)	25.29 (20.01–29.52)	15.99	25.12 (20.58–31.66)	45.99 (35.73–56.71)
OpenPose	9.78	11.64 (10.01–13.55)	20.18 (17.39–23.69)	23.47	37.63 (27.70–45.74)	67.74 (47.80–82.59)
MMPose Wholebody	7.01	6.54 (4.78–8.47)	10.48 (7.81–14.08)	12.28	12.26 (9.10–17.67)	19.58 (14.08–28.58)
OpenPifPaf	6.88	11.96 (8.00–14.38)	22.61 (15.11–26.69)	13.66	24.46 (19.96–28.12)	44.35 (36.52–51.38)
SDPose	13.32	16.53 (11.68–25.10)	29.98 (20.35–45.73)	12.31	12.62 (8.47–17.52)	18.37 (13.62–26.73)
Sapiens	8.27	8.42 (6.39–14.92)	13.61 (10.67–27.18)	10.01	9.94 (7.44–15.42)	15.27 (11.92–26.14)
AlphaPose	12.24	13.53 (7.72–21.19)	21.81 (12.82–36.93)	13.94	16.36 (11.23–19.74)	25.85 (18.36–32.92)
SMPLest-X	**5.53**	**3.87** (3.18–4.51)	**5.85** (4.72–6.80)	**7.36**	**5.56** (4.68–7.02)	**7.40** (6.49–9.34)

Table 7: Temporal stability metrics for hand keypoints on Signsuisse and Phoenix. Format as in Table 3: median (IQR), scaled by 100.

Pose Estimator	Signsuisse			Phoenix		
	E_v	J_{acc} (↓)	J_{jerk} (↓)	E_v	J_{acc} (↓)	J_{jerk} (↓)
MediaPipe	0.31	0.27 (0.24–0.31)	0.44 (0.38–0.49)	1.83	1.65 (1.18–2.85)	2.73 (2.07–4.81)
OpenPose	0.39	0.46 (0.42–0.50)	0.81 (0.73–0.87)	3.07	4.15 (2.47–5.73)	7.16 (4.26–10.32)
MMPose Wholebody	0.44	0.64 (0.58–0.67)	1.14 (1.04–1.19)	1.63	1.23 (1.08–1.50)	1.84 (1.63–2.20)
OpenPifPaf	0.32	0.27 (0.23–0.30)	0.43 (0.38–0.48)	1.64	1.08 (0.87–1.60)	1.52 (1.20–2.40)
SDPose	0.31	0.24 (0.22–0.27)	0.38 (0.34–0.42)	1.59	**0.86** (0.77–1.15)	**1.17** (1.01–1.42)
Sapiens	2.20	3.63 (3.17–4.08)	6.64 (5.76–7.45)	3.92	5.82 (5.23–6.56)	10.29 (9.37–11.60)
AlphaPose	0.42	0.55 (0.48–0.58)	0.95 (0.83–1.01)	1.75	1.41 (1.08–2.37)	2.22 (1.68–3.87)
SMPLest-X	0.69	0.81 (0.65–0.99)	1.38 (1.10–1.69)	**1.37**	1.49 (1.27–1.77)	2.50 (2.09–2.97)

Table 8: Temporal stability metrics for face keypoints on Signsuisse and Phoenix. Format as in Table 3: median (IQR), scaled by 100.

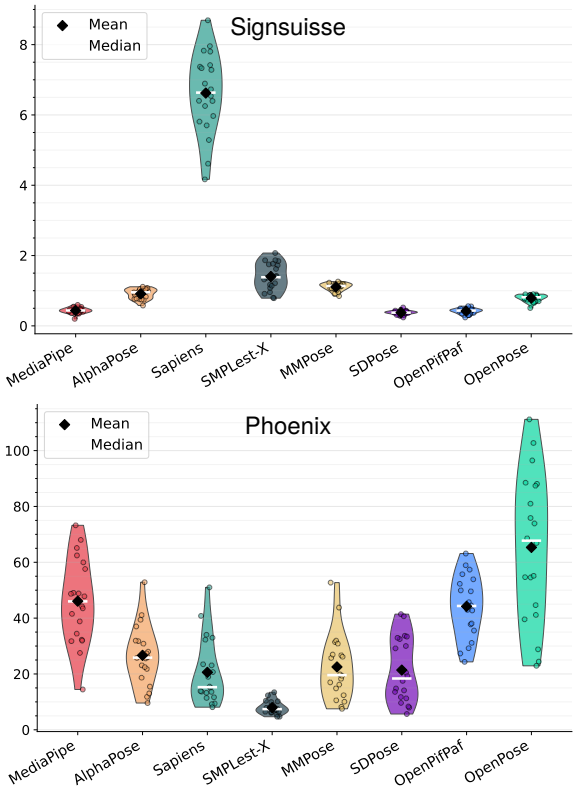


Figure 6: Distribution of per-sequence jerk jitter (J_{jerk}) for face keypoints on Signsuisse (top) and Phoenix (bottom). Format as in Figure 3.

On Signsuisse, SDPose ($J_{\text{jerk}} = 0.38$), MediaPipe (0.44), and OpenPifPaf (0.43) are nearly indistinguishable and achieve the lowest face jitter. Sapiens is a clear outlier (6.64), likely because its dense 243-landmark face mesh amplifies small localization errors into measurable jitter.

On Phoenix, SDPose again achieves the lowest face jitter (1.17), with MMPose Wholebody (1.84) and OpenPifPaf (1.52) also performing well. Sapiens remains the highest (10.29), confirming that its dense face representation is especially sensitive to low-resolution input. MediaPipe shows moderate face jitter (2.73) with a wide IQR, suggesting inconsistent face tracking across sequences.