Meaningful Pose-Based Sign Language Evaluation

Anonymous ACL submission

Abstract

We present a comprehensive study on meaningfully evaluating sign language utterances in the form of human skeletal poses. The study covers keypoint distance-based, embedding-based, and back-translation-based metrics. We show tradeoffs between different metrics in different scenarios through automatic meta-evaluation of sign-level retrieval and a human correlation study of text-to-pose translation across different sign languages. Our findings and the open-source *pose-evaluation* toolkit provide a practical and reproducible way of developing and evaluating sign language translation or generation systems.

1 Introduction

Automatic evaluation metrics are essential for assessing the quality of automatically generated language content and tracking progress over time. For instance, machine translation (MT) studies rely heavily on BLEU (Papineni et al., 2002), even though newer metrics have shown stronger correlation with human judgment (Freitag et al., 2022). This trend continues to sign language processing (SLP; Bragg et al. (2019); Yin et al. (2021)), an interdisciplinary natural language processing and computer vision subfield. Sign language translation (SLT; Müller et al. (2022, 2023a); De Coster et al. (2023)), denoting the part of SLP concerned with translating sign language videos into spoken language text, reuses text-based metrics.

Müller et al. (2023b) puts forward concrete suggestions on evaluating generated text (especially glosses) in a sign language context. They suggest always computing metrics with standardized tools (e.g., SacreBLEU (Post, 2018) for BLEU) and reporting the metric signatures for reproducibility and fair comparison with other work. The opposite direction—generating or translating into sign language utterances (usually from source text)—presents additional challenges for evaluation. Namely, standardized metrics, tooling, and correlation with human evaluation are lacking.

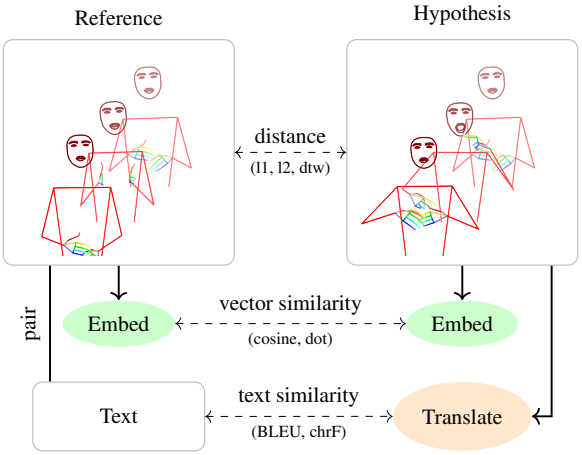


Figure 1: Overview of our pose-based evaluation taxonomy. We compare a reference and a hypothesis pose sequence by (a) computing distance-based metrics directly on the sequences; (b) encoding each sequence into a shared embedding space and measuring similarity; and (c) back-translating the hypothesis poses into text to apply conventional machine translation metrics.

In this work, we systematically examine the metrics employed for evaluating sign language output, especially formatted as human skeletal poses (Zheng et al., 2023) that contain motion of signing (e.g., MediaPipe Holistic; Lugaresi et al. (2019); Grishchenko and Bazarevsky (2020)). We start by a literature review of current research practices in §2, and summarize two major families of metrics: (a) distance-based metrics (§3.1) informed by human motion generation (§2.3) and sign language assessment (SLA; §2.4), assuming the access to references; (b) back-translation-based metrics (§3.3) borrowed from MT (Zhuo et al., 2023) and speech translation (Zhang et al., 2023), assuming the pre-existence of a pose-to-text translation model.

After the initial conceptual review, we select, implement, and meta-evaluate typical metrics along

with additional innovative ones proposed by us (as summarised in Figure 1), through two empirical approaches: automatic meta-evaluation with a sign-level retrieval task (§4); and a sentence-level correlation study between metrics of interest versus deaf evaluator ratings on three text-to-pose MT systems in three spoken-sign language pairs (§5).

We find that keypoint distance-based metrics, when carefully tuned, can rival more advanced approaches for sign retrieval and human-judgment correlation. On the other hand, embedding-based metrics, including ones borrowed from SLA, excel at their own domain but struggle at the sentence level across different systems. Back-translation likelihood emerges as the most consistent metric—highlighting the need for open, standardized pose-to-text models alongside human evaluation.

The source code of the suggested evaluation metrics and the proposed meta-evaluation protocols in §4 are openly maintained in *pose-evaluation*, a public GitHub repository[1]. The human correlation data and evaluation scripts in §5 will also be released to encourage future research.

2 Related Work

We discuss four related fields in this section with a special emphasis on the evaluation methodology, and outline recent work in sign language generation (SLG; §2.2) in Table 1. The remaining three fields provide additional background relevant to evaluating these SLG systems.

2.1 Sign Language Understanding

Sign language recognition (Adaloglou et al., 2021) and translation (De Coster et al., 2023) are the two most prevalent tasks of understanding sign language from video recordings. The former aims at classifying signing into a fixed vocabulary of signs in a particular sign language, either from isolated video clips of single signs (isolated sign language recognition; ISLR) or continuous video footage spanning multiple signs (continuous sign language recognition; CSLR). Given the classification nature, the evaluation efficiently uses classic statistical metrics such as accuracy, F_1 score, and word error rate.

Early SLT attempts rely on glosses (Moryossef et al., 2021b; Müller et al., 2023b), produced manually by humans or a CSLR model. Camgoz et al. (2018, 2020) starts end-to-end neural SLT and leads a wave of gloss-free SLT work (Zhou et al., 2023; Zhang et al., 2024a), where evaluation is typically done with BLEU and BLEURT (Sellam et al., 2020) but not possible with source-based metrics like COMET and quality estimation models like COMET-QE (Chimoto and Bassett, 2022) due to the input modality constraint on sign language. WMT-SLT campaigns for two consecutive years (Müller et al., 2022, 2023a) carry out a rigorous human evaluation process as seen in traditional MT research. Yet the correlation between automatic evaluation metrics and human judgments in SLT has not been reported; quantifying this correlation would yield valuable insights.

2.2 Sign Language Generation

The landscape of SLG is more complicated than SLT, with various inputs, namely, (a) spoken language *text*; (b) sign language *glosses*; (c) iconic *phonetic writing systems* of sign language; (d) textual *phonetic description* of signing, and various outputs, usually, 2D/3D pose; or RGB video frames. We note that in the case of (a) *text*, the generation process involves translation from a spoken language to a sign language with possibly reordering and rephrasing of words, while starting with (b), (c), and (d) merely convert sign language approximated in textual forms into visuals (also known as sign language production[2]), possibly with a preceding step in the pipeline that translates from *text* to (b) (Jiang et al., 2023), (c) (Zhu et al., 2023), and (d). Our work evaluates poses as the primary representation of sign language motion and semantics, deliberately excluding RGB videos to avoid confounding factors such as visual appeal or signer identity. Evaluating videos using the same methods is possible after first estimating them into poses.

We present prominent pose-based SLG studies from recent years and their evaluation methods in Table 1, grouped by the input modalities. Following a similar roadmap as SLT, SLG takes off with a gloss-based cascading approach (text-to-gloss-to-sign; Stoll et al. (2018, 2020)) and then gradually switches to an end-to-end fashion in a series of follow-up work (Saunders et al., 2020b,a, 2021a,b). Attempts have also been made with alternative phonetic inputs such as HamNoSys (Prillwitz and Zienert, 1990; Arkushin et al., 2023).

Unlike SLT, however, gloss-based baseline ap-

[1]The link will be revealed after anonymous review.

[2]The terms are sometimes used interchangeably and thus confuse. This work sticks to the broad term of sign language generation that generates signing from any source.

Work	Datasets	Sources	Target	Model		Evaluation Metrics					
	(P,H2, etc.)	(T,G,H)	(M,O,S)	</>	θ	D	</>	B4	</>	θ	Other
Arkushin et al. (2023)	DGS Corpus, 3 others	H	O	✔	✔	✔	✔	n/a	n/a	n/a	-
Stoll et al. (2018, 2020)	P	G	O	-	n/a	-	-	-	-	-	SSIM, PSNR, MSE (pixel-wise)
Moryossef et al. (2023b)	Signsuisse	G	M	✔	n/a	-	-	-	-	-	-
Zuo et al. (2024b)	P,CSL-Daily	G	S	✔	n/a	✔	-	✔	✔	✔	Frame temporal consistency
Saunders et al. (2020b,a, 2021a,b)	P	T	O	✔	-	-	-	✔	✔	-	-
Hwang et al. (2021, 2023)	P,H2	T	O	✔	-	✔	-	✔	-	-	Fréchet Gesture Distance
Yin et al. (2024)	P	T	S	-	-	✔	-	✔	-	-	-
Fang et al. (2024a,b)	P,H2,4 others	T	O	-	-	✔	-	✔	-	-	SSIM, Hand SSIM, FID, etc.
Yu et al. (2024)	P,H2,4 others	T,G,H	S	✔	-	✔	-	-	-	-	FID, Diversity, MM-Dist, etc.
Baltatzis et al. (2024)	H2	T	S	-	-	✔	-	✔	-	-	FID
Zuo et al. (2024a)	P,H2,CSL-Daily	T	S	-	-	✔	-	✔	-	-	Latency

Table 1: Literature review of recent works on pose-based sign language generation (May 2025). P=`RWTH-PHOENIX-Weather2014T`, H2=`How2Sign`; T=`Text`, G=`Gloss`, H=`HamNoSys`; M=`MediaPipe`, O=`OpenPose`, S=`SMPL-X`; D=`DTW-MJE` (and other distance-based metrics), B4=`BLEU-4` (and other back translation-based metrics); </> and θ represent the availability of source code and model weights for the generation model and the evaluation metrics (including the back translation model if involved), respectively. The check mark symbols (✔) are clickable links in these columns, and *n/a* denotes not-applicable cases, such as model weights for gloss-based systems and back-translation for HamNoSys input. Other image-based metrics are left as less relevant.

proaches for SLG remain competitive and practical choices (Moryossef et al., 2023b; Zuo et al., 2024b) thanks to accessible sign language dictionary resources that allow for straightforward mapping of glosses to sign language pose sequences. Modern end-to-end approaches utilise vector quantization, diffusion models, and LLMs, and the output pose format spans from classic 2D standards such as MediaPipe Holistic and Openpose (Cao et al., 2019) to 3D SMPL-X (Pavlakos et al., 2019).

Popular datasets used in this line of work include RWTH-PHOENIX-Weather 2014T, in German Sign Language (DGS), introduced by Forster et al. (2014); Camgoz et al. (2018); CSL-Daily, in Chinese Sign Language (CSL), introduced by Zhou et al. (2021); and How2Sign, in American Sign Language (ASL), introduced by (Duarte et al., 2021). We choose Signsuisse (Müller et al., 2023a) in this work (§5) for its multilinguality nature and richer vocabulary than others[3].

As for evaluation, the SLRTP Sign Language Production Challenge 2025 summarises the most common evaluation metrics: (a) keypoint distance-based, such as DTW-MJE (Dynamic Time Warping - Mean Joint Error); and (b) back-translation-based, such as BLEU and BLEURT. Human evaluation is conducted briefly in Saunders et al. (2021a,b), Baltatzis et al. (2024), and Zuo et al. (2024a) and more extensively in another concluded campaign–Quality Evaluation of Sign Language Avatars Translation (Yuan et al., 2024). Unfortunately, like in SLT, the correlation between automatic metrics and human judgments has never been formally validated. Upon reviewing Table 1, we spot two significant issues in the current development of SLG: (a) Most systems and their evaluations are non-reproducible due to the lack of source code and model weights (including the back-translation models if involved). (b) Cross-work comparisons are unrealistic given the fragmented implementation of the evaluation metrics (in contrast to MT, where standardized tools like SacreBLEU are available).

2.3 Human Motion Generation

Motion generation from natural language is a related field where human pose sequences are synthesized to reflect described actions (Tevet et al., 2022; Zhang et al., 2024b). Evaluation typically involves distance-based metrics (e.g., joint or velocity error), perceptual similarity (e.g., Fréchet Inception Distance adapted to motion), and alignment metrics like R-Precision to measure text-motion coherence. However, Voas et al. (2023) shows that many of these automated metrics correlate poorly with human judgment on a per-sample basis. They propose MoBERT, a BERT-based learned evaluator, which achieves higher agreement with human ratings, highlighting the ongoing challenge of designing semantically meaningful motion evaluation.

2.4 Sign Language Assessment

SLA research compares student-produced signing against canonical references. Cory et al. (2024) evaluates sign language proficiency by modeling the natural distribution of signing motion and demonstrating a strong correlation with human

[3]PHOENIX and CSL-Daily feature 1066 and 2000 signs.

ratings. Tarigopula et al. (2024, 2025) proposes a posterior-based analysis of skeletal or spatio-temporal features to assess both manual and non-manual signing components, improving alignment with human evaluation.

3 Evaluation Metrics

In this section, we formally define the evaluation metrics mentioned in related work (§2) and implement them, reusing open-source code where available, to prepare for the upcoming empirical study on pose evaluation in §4 and §5.

3.1 Keypoint Distance-Based Metrics

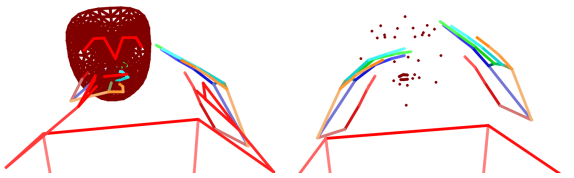
We borrow keypoint distance-based metrics from prior work on sign language generation, notably Ham2Pose (Arkushin et al., 2023). These metrics—originally developed for general pose estimation and motion analysis—quantify geometric similarity using frame-wise errors and alignment strategies. However, they are not designed for sign language and tend to ignore critical linguistic properties such as signer speed variation, hand dominance, and missing keypoints. Moreover, they have not been systematically validated against human judgments in sign language contexts, motivating our extended investigation.

We identify significant sources of variation that affect the outcomes of distance-based metrics: (a) whether and how keypoints are normalized (e.g., based on shoulder positions); (b) whether videos are trimmed to exclude signing-inactive frames; (c) which subset of MediaPipe keypoints is selected for comparison (Figure 2; e.g., hands-only vs. full body); (d) how framerate mismatches are handled (e.g., interpolating to a consistent FPS); (e) how masked or missing keypoints are treated (e.g., filled with a value vs. ignored); and (f) how sequences of unequal length are aligned (Figure 3; using zero-padding, frame repetition, or DTW). We test these variations and provide a reproducible toolkit that enables researchers to tune these design choices explicitly rather than inherit arbitrary defaults.

By mixing preprocessors—including normalization, keypoint selection, masking strategies, trimming, and alignment—with different distance measures, our framework supports the generation of thousands of metric variants to be tested in §4.

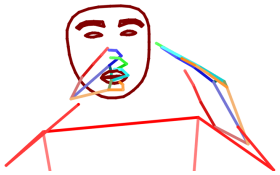
3.2 Embedding-Based Metrics

Rather than operating on the keypoints' raw spatial positions, we categorize embedding-based metrics that calculate distance or similarity in a latent embedding space provided by a model.



(a) Full (586 Points) (Grishchenko and Bazarevsky, 2020)

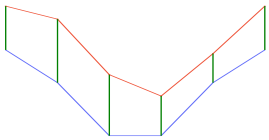
(b) YouTube-ASL (85 Points) (Uthus et al., 2023)



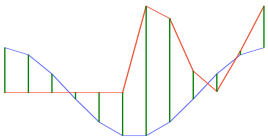
(c) Face-contour (178 Points) (Moryossef et al., 2021a)

(d) SignCLIP (203 Points) (Jiang et al., 2024)

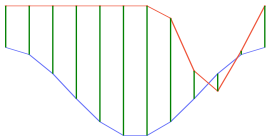
Figure 2: MediaPipe keypoint selection strategies.



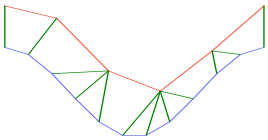
(a) Equal-Length Sequences



(b) Zero-Padding



(c) Pad with First Frame



(d) Dynamic Time Warping

Figure 3: Strategies for sequence alignment between the shorter sequence (in red) and longer sequence (in blue). In reality, pose keypoint trajectories are aligned temporally in 3D and then averaged for the whole body.

3.2.1 Sign Language Assessment Metrics

We adopt two metrics for comparing two poses from the SLA task (§2.4): the Skeleton Variational Autoencoder (SkeletonVAE) model from Cory et al. (2024) and the posterior-based scores from assessment models developed in Tarigopula et al. (2024).

SkeletonVAE Score The SkeletonVAE is trained to produce a per-frame latent embedding. 2D MediaPipe poses are first uplifted to constrained 3D skeletons using the method of Ivashechkin et al. (2023) and then embedded into a 10-dimensional β-VAE latent space (Higgins et al., 2017). We define *SkeletonVAE Score* as the L2 distance between the reference and hypothesis sequences' DTW-aligned latent trajectories, optionally normalized by the DTW path length.

Skeleton Posterior-based SKL Score Following Tarigopula et al. (2024), we first extract two sets of linguistically informed features from the pose sequences with the same missing keypoint preprocessing as Eq. 6 in Arkushin et al. (2023). For hand movement, we compute 36-dimensional feature vectors representing hand position and velocity relative to the head, shoulders, and hips with a temporal context of 9 frames. For handshape, we calculate joint positions relative to the wrist and input them into a separate MLP to obtain handshape posteriors. The resulting stack of shape and movement posteriors from both the reference and hypothesis examples is then aligned using DTW with a cost function based on the Symmetric Kullback–Leibler (SKL) divergence. The cost is aggregated over the DTW time steps as the final score with two variants–*SKL_mvt Score* (movement only) and *SKL_mvt_hshp Score* (movement + handshape), respectively.

3.2.2 SignCLIP Score

One step further than §3.2.1, we follow *CLIPScore* (Hessel et al., 2021) and use SignCLIP (Jiang et al., 2024), a model repurposed for representing sign language poses by multilingual contrastive learning, to derive *SignCLIPScore P-P* (pose-to-pose), based on the dot product of the embeddings of the reference and hypothesis on the sentence level instead of frame-level latents plus DTW alignment.

Reference-Free Quality Estimation Variant We introduce *SignCLIPScore P-T* (pose-to-text). It computes the dot product between the text and pose embedding, eliminating reliance on scarce or even unreliable ground-truth signing references (Freitag et al., 2023).

3.3 Back-Translation-Based Metrics

Assuming the existence of the corresponding spoken language text and a reliable pose-to-text SLT model, we can evaluate a sign language pose by: (a) *Sampling*: translate the pose sequence into text, then compare with the source text using BLEU, chrF, or BLEURT. (b) *Scoring*: compute the log-likelihood of the text given the pose sequence as input to the SLT model. This avoids errors introduced by decoding and supports more consistent comparisons across systems. In this study, we adopt an SLT model from Zhang et al. (2024a), which is pretrained on a large-scale YouTube SLT corpus and massive MT data. We use *system 8* from their study (*YT-Full + Aug-YT-ASL&MT-Large + ByT5 XL*), i.e., the current state of the art, and preprocess the generated pose sequences by selecting the same 85 keypoints specified in their paper[4].

4 Automatic Meta-Evaluation

We adopt the retrieval-based evaluation protocol from Arkushin et al. (2023) to assess how well different metrics capture meaningful distinctions between signs. Each pose sequence is treated as a query, and the goal is to retrieve other samples of the same sign (*targets*) from a pool that includes unrelated signs (*distractors*). We focus primarily on the distance-based metric variants introduced in §3.1, and compare them against embedding-based alternatives such as *SignCLIP Score* (§3.2.2). The key results are presented in Table 2.

Base	Def	Fill	Trim	Norm.	Pad	Keypoints	**mAP↑**	**P@10↑**
APE	10	10	x	x	zero	Up. Body	33%	27%
APE	10	10	x	x	first	Up. Body	34%	29%
APE	10	10	v	x	zero	Up. Body	35%	30%
APE	10	10	x	v	zero	Reduced	36%	32%
APE	10	10	x	x	first	Reduced	37%	32%
APE	10	10	x	x	first	YT-ASL	39%	36%
APE	10	10	v	x	first	Reduced	40%	36%
APE	10	10	v	v	zero	Up. Body	41%	37%
APE	10	10	x	x	zero	Hands	42%	38%
APE	10	10	v	v	first	YT-ASL	43%	39%
APE	10	10	v	x	zero	Hands	45%	41%
DTW	10	10	x	x	/	Up. Body	36%	32%
DTW	10	10	v	x	/	Up. Body	37%	33%
DTW	10	10	x	x	/	Reduced	42%	40%
DTW	10	10	x	v	/	Up. Body	43%	41%
DTW	10	10	v	v	/	Up. Body	43%	41%
DTW	10	10	x	v	/	Reduced	44%	41%
DTW	10	10	x	v	/	Hands	45%	41%
DTW	10	10	x	x	/	YT-ASL	48%	47%
DTW	10	10	v	x	/	YT-ASL	49%	48%
DTW	10	10	x	x	/	Hands	53%	52%
DTW‡	0	10	v	x	/	Hands	53%	52%
DTW	10	10	v	x	/	Hands	53%	52%
DTW†	1	1	x	v	/	Hands	55%	53%
SignCLIPScore P-P (multilingual)							50%	48%
SignCLIPScore P-P (ASL Finetuned)							**91%**	**92%**

Table 2: Automatic meta-evaluation of reference-based metrics on retrieval. A representative sample, including the best-performing configurations, is shown in this table. "*Def*" indicates default distance; "*Fill*" indicates the value used to fill in missing keypoint; "*zero*" indicates zero-padding; "*first*" indicates padding with first frame.

Evaluation is conducted on a combined dataset of ASL Citizen (Desai et al., 2023), Sem-Lex (Kezar et al., 2023), and PopSign ASL (Starner et al., 2023). For each sign class, we use all available samples as targets and sample four times as many distractors, yielding a 1:4 target-to-distractor

[4]Eight mismatched keypoints due to different MediaPipe versions are imputed as missing landmarks.

ratio. For instance, for the sign *HOUSE* with 40 samples (11 from ASL Citizen, 29 from Sem-Lex), we add 160 distractors and compute pairwise metrics from each target to all 199 other examples[5]. Retrieval quality is measured using Mean Average Precision (mAP↑) and Precision@10 (P@10↑). The complete evaluation covers 5362 unique sign classes and 82,099 pose sequences.

After several pilot runs, we finalized a subset of 169 sign classes with at most 20 samples each, ensuring consistent metric coverage. We evaluated 48 distance-based variants and SignCLIP models with different checkpoints provided by the authors on this subset. The overall results show that DTW-based metrics outperform padding-based baselines. Embedding-based methods, particularly SignCLIP models fine-tuned on in-domain ASL data, achieve the strongest retrieval scores.

5 Text-to-Pose Translation Study with Human Evaluation

This section shifts our evaluation focus from automatic sign-level tasks to a sentence-level text-to-pose sign language machine translation scenario. Due to their subjective and diverse nature, open-ended text or utterance generation tasks inherently lack a single "correct"/"ground-truth" answer. Consequently, **automatic evaluation metrics are only *meaningful* if they correlate closely with human judgments** (Reiter, 2018; Sellam et al., 2020).

5.1 Dataset: WMT-SLT Signsuisse

We use the Signsuisse dataset released in the WMT-SLT campaign. The dataset comprises 18,221 lexical items in three spoken-sign language pairs, represented as videos and glosses. One signed example sentence for each lexical item is presented in a video along with the corresponding spoken language translation, which forms parallel data between the sign and spoken languages. The test set is used to test different text-to-pose translation systems. It contains 500 German/Swiss German Sign Language (DSGS) segments, 250 French/French Sign Language (LSF) segments, and 250 Italian/Italian Sign Language (LIS) segments.

5.2 Systems

We utilize three text-to-pose translation systems that convert spoken language text inputs into corresponding sign language represented by the MediaPipe Holistic pose formats.

Reference* MediaPipe poses are estimated from the reference translation videos.

sign.mt Based on Moryossef et al. (2023b), this open system converts text into sign language glosses through rule-based reordering and selective word dropping. Glosses are mapped to skeletal poses retrieved from a lexicon and are then concatenated to form coherent sequences. When a gloss is missing from the lexicon, the system defaults to fingerspelling the corresponding word.

sign.mt v2 During evaluation, we found that frequent fingerspelling of missing glosses was cumbersome and frustrating for evaluators. Therefore, in this version, we opted to omit glosses without lexical mappings, acknowledging that while this may result in information loss, it significantly improves user experience and evaluation efficiency.

Sockeye We adapt Sockeye (Hieber et al., 2022) to continuous pose sequences by modifying both the encoder and decoder to handle continuous sequences. The text-to-pose Sockeye model is trained on the Signsuisse training set with 60k updates on a 32GB NVIDIA Tesla V100 GPU.

To avoid exposure bias—where the decoder overfits to gold frames and fails at inference, we first predict only the initial pose y_1 from the encoder output, then feed y_1 as input for all subsequent steps $y_{2:n}$, training the decoder to output frame-to-frame deltas $\Delta y_t = y_t - y_1$ instead of absolute poses. Since the target sequence is continuous, we replace the cross-entropy loss function with mean squared error on the poses. Additionally, there is no <EOS> token with continuous output; instead, we learn to output the length of the pose sequence based on the length ratios from the training data[6].

5.3 Human Evaluation

We collect system translations and use Appraise (Federmann (2018); Figure 4) to allow evaluators to rate the translations on a continuous scale between 0 and 100 as in traditional direct assessment (Graham et al., 2013; Cettolo et al., 2017) but with 0-6 markings on the analogue slider and custom annotator guidelines designed explicitly for our task (similar to WMT-SLT, but reverse translation direction). Evaluation instructions are sent out in DSGS,

[5]We consistently discard scores for pose files where either the target or distractor could not be embedded with SignCLIP.

[6]Public repo with demo outputs, link hidden for anonymity.

	Reference-Based							Reference-Free						
	Distance-Based				SLA Metrics			SignCLIPScore		Back Translation-Based				H*
	nAPE	**nDTW**	**DTWp**	**nDTWp**	**SVAE**	**SVAE$_n$**	**SKL**	**P-P**	**P-T**	**B4**	**chrF**	**B-RT**	**Lik.**	**H***
By System														
sign.mt	0.09	0.14	0.11	0.10	0.23	-0.08	0.24	0.10	0.02	0.05	0.11	0.05	0.23	0.43
sign.mt v2	0.28	0.33	0.26	0.31	0.46	0.14	0.22	0.00	-0.19	0.20	0.22	0.44	0.49	0.52
Sockeye	0.10	0.15	0.04	0.17	0.13	0.01	0.24	0.42	-0.27	-0.07	0.04	0.46	0.58	0.22
By Language														
DE→DSGS	-0.36	-0.09	0.73	0.43	-0.02	0.27	-0.57	-0.31	0.39	0.18	0.26	0.09	0.36	0.70
FR→LSF	-0.54	-0.11	0.76	0.02	-0.01	0.37	-0.68	-0.01	0.45	0.32	0.60	0.47	0.29	0.80
IT→LIS	-0.57	-0.39	0.79	0.57	-0.02	0.53	-0.75	0.13	0.29	0.31	0.63	0.41	0.38	0.88
Overall (↑)	-0.41	-0.10	0.76	0.43	0.07	0.38	-0.56	-0.10	0.27	0.21	0.42	0.36	0.42	0.77
SD (↓)	(0.35)	(0.24)	(0.34)	(0.20)	(0.18)	(0.22)	(0.47)	(0.22)	(0.29)	(0.14)	(0.23)	(0.18)	(0.12)	(0.24)

Table 3: Segment-level Spearman correlations with average human judgments calculated for several pose-based evaluation metrics for sign language. nAPE=`normalized APE`, nDTW=`normalized DTW-MJE` (two metrics taken from Arkushin et al. (2023) and re-implemented for MediaPipe, normalized by pose shoulder); DTWp=`DTW+Trim+Default0.0+Hands-Only`, nDTWp=`DTW+Default1.0+MaskFill1.0+Norm.+Hands-Only` (top metrics selected in §4 implemented by *pose-evaluation*, denotated by ‡ and † in Table 2, without/with pose normalization, respectively); SVAE=`SkeletonVAE Score`, SVAE$_n$=`SVAE normalized by DTW path`, SKL=`SKL_mvt Score`; P-P=`Pose-to-pose embedding distance`, P-T=`Pose-to-text embedding distance`; B4=`BLEU-4`, chrF=`chrF`, B-RT=`BLEURT`, Lik. =`Likelihood`. H* denotes mean inter-evaluator Spearman correlation. SD represents the standard deviation across each column and is expected to be small/consistent for an ideal metric.

LSF, and LIS, which are translations of the respective spoken language instructions in WMT-SLT.

We hire seven DSGS, two LSF, and four LIS evaluators, all of whom are native deaf sign language users[7]. All work is done with informed consent in written and signed form. Of the seven native DSGS deaf signers, four have never participated in such an evaluation campaign before, two have done so once, and one has already attended more than once. Concerning their professional backgrounds, four are deaf translators; one also interprets live. Complete demographics are presented in Table 4.

An initial round of evaluation informs us about the cost, roughly 100 example segments per hour, with a compensation of ∼40 USD per hour. Evaluators also provide constructive feedback on the Appraise platform and the translation systems, which results in the v2 version of sign.mt. Therefore, the number of evaluated examples varies between systems and languages.

Statistics The evaluation comprises 11,471 ratings of 2650 unique examples across all four (three plus reference) systems and three language pairs. We follow the practices set by WMT-SLT. The inter-annotator agreement, measured with an approximation of Fleiss κ (Fleiss, 1971) by discretizing the continuous scale 0-100 in seven bins in the scale 0-6, is $\kappa = 0.36 \pm 0.05$. We also randomly mix 500 references and some repeated hypothesis segments for sanity checks and quality control. The mean intra-annotator agreement over all evaluators is $\kappa = 0.49 \pm 0.09$, calculated over 50-100 segments evaluated twice by the same evaluator. We find the inter- and intra-annotator agreement lower than in the WMT-SLT study for the reverse translation direction and posit the lack of a clear definition and criteria of translation quality on signing poses.

[7]One additional DSGS evaluator, a hearing interpreter, did a pilot study with us to test the Appraise system.

	Evaluation experience			SL professional			Avg yr.
	Never	Once	> Once	Translator	Interpreter	Teacher	
DSGS (7)	4	2	1	4	1	4	39.0
LSF (2)	0	1	1	1	0	1	35.0
LIS (4)	1	1	2	3	4	0	42.5

Table 4: Raters overview: system evaluation and professional experience with sign language, average number of years signing (in most cases equivalent to age).

5.4 Correlation Analysis

We run a correlation analysis between the metrics proposed in §3 divided into different families and the human scores averaged over evaluators on the segment level, as presented in Table 3. The absolute scores per metric/system are appended in Table 5.

For the distance-based metrics, we reproduce *nAPE* and *nDTW* for MediaPipe poses based on the open implementation from Arkushin et al. (2023) as a reference, and additionally compare them to

the best-performing metrics informed by the automatic meta-evaluation in §4 on ASL, a different sign language. We flip the signs of the metrics that quantify errors to keep a positive correlation for analytical convenience. Row-wise, we first break down the correlation by systems and languages into relevant rows, and finally present the overall correlation, including all systems and languages, to also reflect the performance on the system level.

6 Discussion and Recommendations

Distance-based metrics are efficient defaults, but the devil is in the implementation details. Although seemingly straightforward to implement, distance-based metrics involve many details concerning the pose format, keypoint selection, and other corner cases. We empirically show the effectiveness of correcting these design choices by a random parameter search, following our meta-evaluation protocols established in §4. We recommend using the tuned version– *DTWp* and *nDTWp*– in our *pose-evaluation* library, or tuning your distance-based metrics if there is a special scenario.

Upon successful tuning, a distance-based metric achieves decent sign retrieval and correlation with humans in the text-to-pose translation. Our tuned metrics can be used as a distance function for a nearest neighbor classifier, and reach close performance as the SignCLIP model pretrained on multilingual sign language data; still, it lags behind a SignCLIP model fine-tuned on in-domain data (Table 2). When used to evaluate translation output, keypoint distance-based metrics can range from negatively correlated with human judgments (as seen for *nAPE* and *nDTW*), to being the best metrics tested. *DTWp* wins the overall correlation while *nDTWp* is more sensitive on the segment level within a specific system (Table 3).

SLA metrics correlate with humans on the segment level, but are confused on the system level. While verified to align with human ratings for their tasks on evaluating human-produced signing (usually fixed individual signs), text-to-pose translation is more lengthy and open-ended, which hurts the direct transferability. A proper length normalization (as seen in the case of $SVAE_n$ vs. *SVAE*) might help on the system level at the price of losing precision on the segment level.

SignCLIP, used as a multilingual embedding device, excels on the sign level, but falls short for sentence-level translation evaluation. We speculate that using a single embedding to summarize a long-duration ($>$ 10 seconds) signing video is inherently limited, especially for DSGS, an unseen language during SignCLIP pretraining. Nevertheless, the reference-free variant shows moderate correlation on the system level, and we observe a similar tradeoff (*P-P* vs. *P-T*) between segment and system level correlations as observed in the SLA metrics.

Back-translation-based approaches correlate properly with human judgment; a gap remains compared to inter-human correlation. In addition to the standard practices suggested by Müller et al. (2023b) on computing text-based metrics, we call for open, standardized pose-to-text translation models that include both the model weights and the source code. Yet, as put by Table 1, it is hardly the case in current research, and having one dedicated back translation model for each translation direction (or even dataset) is a luxury. The above-mentioned metrics, which do not rely on in-domain data but function to a decent degree, are valuable in a more generic setting. Human evaluation shall be used as the final quality assessment resort.

When using back translation, likelihood is consistent and more reliable than text metrics. BLEU, chrF, and BLEURT show weaker or unstable correlations with humans in Table 3. It is recommended that back-translation likelihood be included as a primary metric, when a pose-to-text model is available.

7 Conclusion

This work presents a unified framework and an open-source *pose-evaluation* toolkit for systematically assessing (generated) sign language utterances based on human skeletal poses. We implemented and compared a wide range of metrics (§3)—distance-based, embedding-based, and back-translation-based—via automatic meta-evaluation on sign retrieval (§4) and a comprehensive human correlation study across three sign languages (§5). Our results demonstrate that carefully tuned distance metrics, namely *DTWp* and *nDTWp*, and back-translation likelihoods yield the strongest agreement with native signer judgments. We release our code, evaluation protocols, and human ratings to foster reproducible and fair comparisons in computational sign language research.

8 Limitations

8.1 3D Pose Representation

While our study focuses on using MediaPipe Holistic as the pose format for representing sign language motion, other specifics, especially the recently developed 3D SMPL-X (Pavlakos et al., 2019) would be a visually more expressive choice. However, the lack of a common way to extract and use 3D poses as easily as MediaPipe Holistic makes the latter the most used choice in SLP.

8.2 Missing Publicly Available Systems

Our study is further limited by the number of public systems (Table 1) we can use to run the correlation analysis, unless we implement everything from scratch (including the pose estimation pipelines, text-to-pose systems, and back-translation models). We hope the release of this work will alleviate the situation.

8.3 Automatic Evaluation beyond Sign Level

The automatic meta-evaluation in §4 is capped by the sign-level retrieval task, and we envision extending it to phrase-level. One possible approach is to leverage the Platonic Representation Hypothesis proposed by Huh et al. (2024). In the pose evaluation scenario, we hypothesize that the similarity given by a good pose metric between two pose segments should correlate with the similarity given by a text embedding model between the two text segments paired with the two pose segments, respectively. We leave exploration on this end to future work, which will likely connect more closely the automatic meta-evaluation to the sentence-level human correlation study in §5.

8.4 Tokenized Evaluation

Inspired by how text metrics like BLEU collect surface-form overlapping statistics, we imagine a tokenized evaluation to be promising for sign language evaluation. Although a sign language pose sequence cannot be discretely tokenized and matched like text tokens, the combination of a sign language segmentation model (Moryossef et al., 2023a) plus SignCLIP embedding can be utilized in a way similar to BERTScore (Zhang* et al., 2020), where a similarity matrix is constructed between the reference and hypothesis tokens to derive the final similarity score on phrase level.

References

Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. 2021. A comprehensive study on deep learning-based methods for sign language recognition. *IEEE Transactions on Multimedia*, 24:1750–1762.

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

Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. 2024. Neural sign actors: A diffusion model for 3d sign language production from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033.

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*.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Niehues Jan, Stüker Sebastian, Sudoh Katsuitho, Yoshino Koichiro, and Federmann Christian. 2017. Overview of the IWSLT 2017 evaluation campaign. In *International Workshop on Spoken Language Translation*, pages 2–14.

Everlyn Asiko Chimoto and Bruce A. Bassett. 2022. COMET-QE and active learning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4735–4740, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Oliver Cory, Ozge Mercanoglu Sincan, Matthew Vowels, Alessia Battisti, Franz Holzknecht, Katja Tissi, Sandra Sidler-Miserez, Tobias Haug, Sarah Ebling, and Richard Bowden. 2024. Modelling the distribution of human motion for sign language assessment. In *Proceedings of the 12th Workshop on Assistive Computer Vision and Robotics (ACVR) at ECCV*.

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2023. Machine translation from signed to spoken languages: State of the art and challenges. *Universal Access in the Information Society*, pages 1–27.

Aashaka Desai, Lauren Berger, Fyodor O. Minakov, Vanessa Milan, Chinmay Singh, Kriston Pumphrey, Richard E. Ladner, Hal Daum'e, Alex X. Lu, Naomi K. Caselli, and Danielle Bragg. 2023. Asl citizen: A community-sourced dataset for advancing isolated sign language recognition. *ArXiv*, abs/2304.05934.

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sen Fang, Chunyu Sui, Yanghao Zhou, Xuedong Zhang, Hongbin Zhong, Minyu Zhao, Yapeng Tian, and Chen Chen. 2024a. Signdiff: Diffusion models for american sign language production. *Preprint*, arXiv:2308.16082.

Sen Fang, Lei Wang, Ce Zheng, Yapeng Tian, and Chen Chen. 2024b. Signllm: Sign languages production large language models. *Preprint*, arXiv:2405.10718.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological bulletin*, 76(5):378.

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).

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Ivan Grishchenko and Valentin Bazarevsky. 2020. Mediapipe holistic.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*.

Felix Hieber, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, et al. 2022. Sockeye 3: Fast neural machine translation with pytorch. *arXiv preprint arXiv:2207.05851*.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20617–20642. PMLR.

Eui Jun Hwang, Jung-Ho Kim, and Jong C Park. 2021. Non-autoregressive sign language production with gaussian space. In *BMVC*, volume 1, page 3.

Eui Jun Hwang, Huije Lee, and Jong C Park. 2023. Autoregressive sign language production: A gloss-free approach with discrete representations. *arXiv preprint arXiv:2309.12179*.

Maksym Ivashechkin, Oscar Mendez, and Richard Bowden. 2023. Improving 3d pose estimation for sign language. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5.

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.

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, Miami, Florida, USA. Association for Computational Linguistics.

Lee Kezar, Elana Pontecorvo, Adele Daniels, Connor Baer, Ruth Ferster, Lauren Berger, Jesse Thomason, Zed Sevcikova Sehyr, and Naomi Caselli. 2023. The sem-lex benchmark: Modeling asl signs and their phonemes. *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Amit Moryossef, Zifan Jiang, Mathias Müller, Sarah Ebling, and Yoav Goldberg. 2023a. Linguistically motivated sign language segmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12703–12724, Singapore. Association for Computational Linguistics.

Amit Moryossef, Mathias Müller, and Rebecka Fahrni. 2021a. pose-format: Library for viewing, augmenting, and handling .pose files. `https://github.com/sign-language-processing/pose`.

Amit Moryossef, Mathias Müller, Anne Göhring, Zifan Jiang, Yoav Goldberg, and Sarah Ebling. 2023b. 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, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021b. Data augmentation for sign language gloss translation. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 1–11, Virtual. Association for Machine Translation in the Americas.

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.

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.

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

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, Belgium, Brussels. 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.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020a. Adversarial Training for Multi-Channel Sign Language Production. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020b. Progressive Transformers for End-to-End Sign Language Production. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021a. Continuous 3D Multi-Channel Sign Language Production via Progressive Transformers and Mixture Density Networks. In *International Journal of Computer Vision (IJCV)*.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021b. Mixed signals: Sign language production via a mixture of motion primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1919–1929.

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.

Thad Starner, Sean Forbes, Matthew So, David Martin, Rohit Sridhar, Gururaj Deshpande, Sam S. Sepah, Sahir Shahryar, Khushi Bhardwaj, Tyler Kwok, Daksh Sehgal, Saad Hassan, Bill Neubauer, Sofia Anandi Vempala, Alec Tan, Jocelyn Heath, Unnathi Kumar, Priyanka Mosur, Tavenner Hall, Rajandeep Singh, Christopher Cui, Glenn Cameron, Sohier Dane, and Garrett Tanzer. 2023. Popsign asl v1.0: An isolated american sign language dataset collected via smartphones. In *Neural Information Processing Systems*.

Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2018. Sign language production using neural machine translation and generative adversarial networks. In *BMVC*, volume 2019, pages 1–12.

Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. Text2sign: towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 128(4):891–908.

Neha Tarigopula, Preyas Garg, Skanda Muralidhar, Sandrine Tornay, Dinesh Babu Jayagopi, and Mathew Magimai.-Doss. 2024. Content-based objective evaluation of artificially generated sign language videos. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3815–3819.

Neha Tarigopula, Sandrine Tornay, Ozge Mercanoglu Sincan, Richard Bowden, and Mathew Magimai Doss. 2025. Posterior-based analysis of spatio-temporal features for sign language assessment. *IEEE Open Journal of Signal Processing*.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. 2022. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 358–374. Springer.

Dave Uthus, Garrett Tanzer, and Manfred Georg. 2023. YouTube-ASL: A Large-Scale, Open-Domain American Sign Language-English Parallel Corpus. *Advances in Neural Information Processing Systems*, 36:29029–29047.

Jordan Voas, Yili Wang, Qixing Huang, and Raymond Mooney. 2023. What is the best automated metric for text to motion generation? In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11.

Aoxiong Yin, Haoyuan Li, Kai Shen, Siliang Tang, and Yueting Zhuang. 2024. T2S-GPT: Dynamic vector quantization for autoregressive sign language production from text. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3345–3356, Bangkok, Thailand. Association for Computational Linguistics.

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.

Zhengdi Yu, Shaoli Huang, Yongkang Cheng, and Tolga Birdal. 2024. Signavatars: A large-scale 3d sign language holistic motion dataset and benchmark. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–19.

Zhao Yuan, Zhang Ruiquan, Yao Dengfeng, and Chen Yidong. 2024. Translation quality evaluation of sign language avatar. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 405–415, Taiyuan, China. Chinese Information Processing Society of China.

Biao Zhang, Garrett Tanzer, and Orhan Firat. 2024a. Scaling sign language translation. *arXiv preprint arXiv:2407.11855*.

Dong Zhang, Rong Ye, Tom Ko, Mingxuan Wang, and Yaqian Zhou. 2023. DUB: Discrete unit back-translation for speech translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7147–7164, Toronto, Canada. Association for Computational Linguistics.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2024b. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4115–4128.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. 2023. Deep learning-based human pose estimation: A survey. *ACM Comput. Surv.*

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20871–20881.

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

Dele Zhu, Vera Czehmann, and Eleftherios Avramidis. 2023. Neural machine translation methods for translating text to sign language glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12523–12541, Toronto, Canada. Association for Computational Linguistics.

Terry Yue Zhuo, Qiongkai Xu, Xuanli He, and Trevor Cohn. 2023. Rethinking round-trip translation for machine translation evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 319–337, Toronto, Canada. Association for Computational Linguistics.

Ronglai Zuo, Rolandos Alexandros Potamias, Evangelos Ververas, Jiankang Deng, and Stefanos Zafeiriou. 2024a. Signs as tokens: An autoregressive multilingual sign language generator. *arXiv preprint arXiv:2411.17799*.

Ronglai Zuo, Fangyun Wei, Zenggui Chen, Brian Mak, Jiaolong Yang, and Xin Tong. 2024b. A simple baseline for spoken language to sign language translation with 3d avatars. In *European Conference on Computer Vision*, pages 36–54. Springer.

A Extended Human Evaluation Details

Figure 4 presents a screenshot of the Appraise platform we customized for the text-to-pose evaluation.

B Extended Text-to-Pose Evaluation

Table 5 presents the mean absolute scores for each metric across different systems, in addition to the correlation analysis in Table 3.

	nAPE↓	**nDTW**↓	**DTW**p↓	**nDTW**p↓	**SVAE**↓	**SVAE**$_n$↓	**SKL**↓	**P-P**↑	**P-T**↑	**B4**↑	**chrF**↑	**B-RT**↑	**Lik.**↑	**H***
reference*	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	74.23	15.05	38.52	0.49	-32.87	76.55
sign.mt	0.60	25.43	5171.55	8.66	1.20	0.0028	2794.29	81.58	75.55	3.46	14.53	0.26	-39.30	22.00
sign.mt v2	1.65	20.73	4508.62	8.97	1.23	0.0045	4165.78	89.89	76.46	6.89	24.79	0.23	-67.55	30.12
Sockeye	0.29	16.97	11879.58	10.71	1.16	0.0057	1056.29	94.45	72.40	3.44	11.90	0.17	-77.57	5.05

Table 5: Mean absolute scores for each metric across systems. Rows and columns mirror those in Table 3.

Figure 4: A screenshot of an example text-to-pose evaluation task in Appraise featuring sentence-level source-based direct assessment custom annotator guidelines in German/French/Italian and DSGS/LSF/LIS, translated into English for readers' convenience.