BackTranslation2.0 - A Linguistically Motivated Metric to Assess Sign Language Production

Oliver Cory, Maksym Ivashechkin, Karahan Sahin, Oline Ranum, Jianhe Low, Edward Fish, Anton Pelykh, Ozge Mercanoglu Sincan, and Richard Bowden

Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK
{o.cory,m.ivashechkin,k.sahin,o.ranum,jianhe.low,edward.fish,a.pelykh,
o.mercanoglusincan,r.bowden}@surrey.ac.uk
https://cogvis-cvssp.github.io/BackTranslation2/

Abstract. Sign Languages (SLs) are the primary means of communication for millions of deaf[1] individuals, yet existing evaluation metrics for generated SL remain simplistic and poorly aligned with human judgements. We introduce BackTranslation2.0, a linguistically grounded evaluation metric for text-to-sign translation that moves beyond naïve backtranslation. Our approach adopts an agentic framework in which a deterministic pipeline orchestrates a suite of specialised tools to assess four scoring dimensions—grammatical correctness, phonological accuracy, motion fluency, and generation fidelity—aligned with human rater assessments. Tool outputs are not treated independently: a set of large language model (LLM)-based cross-referential comparison modules evaluates consistency across tools and checks outputs against linguistic expectations, enabling structured reasoning over grammatical, phonological, and motion-level evidence. Final dimension scores are computed through deterministic weighted formulas over validated tool outputs. To validate BackTranslation2.0, we introduce and evaluate on a British Sign Language (BSL) dataset rated in a human rater study across the same quality dimensions, following a protocol developed in collaboration between linguists and deaf experts, benchmarking against six baseline metrics. Our method demonstrates strong correlation with human judgements across all dimensions, providing a more comprehensive, interpretable, and linguistically principled evaluation framework for sign language production systems.

Keywords: Sign Language Assessment · Agentic Evaluation

1 Introduction

Sign Languages (SLs) are rich visual–gestural languages used by millions of deaf individuals worldwide as their primary means of communication. Unlike spoken

[1] We follow the recent convention of abandoning a distinction between Deaf and deaf and use the latter term also to refer to (deaf) members of the sign language community [25, 32].

languages, SLs convey meaning through the visual modality, combining coordinated manual articulations (handshape, movement, location, and orientation) with non-manual features (NMF) such as facial expression and body posture. Computational approaches to SL processing have been studied for over two decades [33, 53, 56], with rapid progress recently across recognition [6, 28, 50, 69], translation [7, 12, 17, 63], and production [3, 39, 54, 62, 68] tasks.

Sign Language Production (SLP), the generation of SL output from spoken language text, represents a critical frontier for accessibility technology. However, a fundamental challenge remains: *How do we reliably evaluate the quality of generated SL?* Existing automatic evaluation paradigms predominantly rely on backtranslation [45], where generated sign sequences are translated back into text and compared against the source sentence using metrics such as BLEU-4 [34]. While computationally convenient, such approaches correlate poorly with judgements from native deaf raters [20, 44], who evaluate a far broader set of linguistic and visual properties. What's more, translation from sign to text remains unsolved for anything other than simple datasets, so using such approaches means using poorly performing recognition/translation technology to assess the performance of developing production technology. Therefore, the only reliable indicator of production or translation quality is human evaluation.

Human evaluation of SLP assesses multiple linguistic dimensions that determine whether a sequence forms a coherent and interpretable sign expression. Raters examine surface-level linguistic realisations affecting phonological and morphological components [5, 42] such as handshape accuracy, articulation location, and use of signing space and directionality, affecting naturalness and fluency of the signing, while evaluating syntactic and semantic information, such as appropriate word order and non-manual usage in relation to understandability [15, 23, 65]. Studies of human assessment therefore adopt structured, multi-faceted evaluation frameworks to capture these dimensions [1, 2, 36]. However, expert annotation is expensive and time-intensive, limiting scalability [11].

In this work, we introduce **BackTranslation2.0**, a linguistically motivated metric designed to bridge the gap between automatic evaluation and human judgement for text-to-sign translation systems. The metric evaluates four scoring dimensions aligned with five aspects rated by human evaluators: **(i) Grammatical Correctness**, adherence to target SL syntax, including use of signing space, pronominal indexing, and directional verb agreement (directionality); **(ii) Phonological Accuracy**, correctness of sub-lexical components, including handshape, movement, location, orientation, and non-manual features; **(iii) Motion Fluency**, naturalness and temporal smoothness of signing motion; and **(iv) Generation Fidelity**, visual quality and anatomical plausibility of the generated output. We realise this multi-dimensional evaluation through an agentic framework in which a reasoning large language model (LLM) orchestrates a suite of specialised linguistic tools and structured lexical resources. Rather than treating tool outputs independently, the system performs cross-referential comparison: LLM-based modules evaluate consistency across tool outputs and assess them against linguistic expectations, enabling structured reasoning over

grammatical, phonological, and motion-level evidence. Final dimension scores are computed using the weighted combination over validated tool outputs, yielding interpretable and reproducible evaluation. Our contributions are as follows. **(1)** We present BackTranslation2.0, a comprehensive linguistically motivated metric for evaluating text-to-sign translation that is superior to naïve back-translation. **(2)** We propose, to our knowledge, the first agentic framework for SL evaluation, in which specialised tools report structured evidence, the language model returns cross-referential comparison labels, and a fixed deterministic formula aggregates these into a reproducible score. **(3)** We introduce a specialised British Sign Language (BSL) evaluation dataset rated by deaf and hearing signers across multiple quality dimensions. **(4)** We benchmark BackTranslation2.0 against six baseline metrics, demonstrating strong correlation with human judgements across all dimensions and further generalisation at scale, providing a more interpretable, linguistically grounded paradigm for SLP.

2 Related Works

2.1 Automatic Evaluation of Sign Language Production

SLP is the task of generating sign language output from spoken-language text. Depending on the system, this output can take several forms: RGB video generated directly in pixel space [39, 46], articulated 3D body meshes (*e.g.* SMPL-X [35]) [3, 68], pose or skeleton sequences [45, 55], or hybrid representations that combine explicit geometry with neural rendering, such as Gaussian splatting [16, 22]. Yet automatic evaluation remains limited and inconsistent.

One family of metrics is *reference-free* with respect to motion, avoiding the need for paired ground-truth signing. The most common example is backtranslation [45], where the generated output is passed through a sign-to-text model and compared against the source sentence using text-based metrics such as BLEU-4 [34]. Joint-embedding approaches such as SignCLIP [21] follow a similar principle in a pose-to-text (p2t) configuration. Although both avoid a direct motion reference, they remain insensitive to important aspects of sign quality, including acceptable articulatory variation, grammar, spatial structure, and phonological well-formedness.

A second family is *reference-based* with respect to motion, directly comparing a generated production against one or more target sequences. In pose-based settings, Dynamic Time Warping (DTW) is commonly used to align motion trajectories under temporal mismatch and measure low-level similarity [57]. Other approaches compare sequences in a learned latent space, such as the Skeleton-VAE (SVAE) [10, 20], or in shared embedding spaces [21] in a pose-to-pose (p2p) configuration, while SiBLEU [67] scores similarity over discretised sign representations. Although more tolerant of surface variation than direct trajectory matching, these methods all compare against a particular target realisation and therefore cannot fully account for the grammatical, lexical, and articulatory variation present in human signing. Modelling the distribution of acceptable signing

rather than a single canonical trajectory [10] mitigates this, but in practice requires multiple expert renditions of each sentence to estimate that distribution. Taken together, these limitations motivate a metric that evaluates sign production across multiple complementary linguistic and visual dimensions.

Distinct from these metrics, the SLP Tier framework [4] is not an evaluation technique but a taxonomy, developed with the deaf community, that classifies spoken-to-sign systems by level of automation. Analogous to the levels used for autonomous driving, it defines eight tiers, from fully manual translation (Tier 0) to fully automatic translation requiring no human intervention (Tier 7). Its value here is in defining *what matters* for a competent system through five core attributes: photo-realistic appearance, SL grammar and word order, correct use of space and direction, non-manual features, and fluid, natural motion, all required to reach the highest partially automated tier. The taxonomy is explicitly not a substitute for accuracy and is meant to sit alongside a metric that quantifies understandability. BackTranslation2.0 is that metric: its dimensions—grammatical correctness (with spatial and directional use), phonological accuracy (with non-manual features), motion fluency, and generation fidelity—target precisely these categories, measuring each directly and aligning evaluation with the priorities of deaf experts and the wider deaf community.

2.2 LLMs as Evaluators and Tool-Using Agents

LLMs perform well not only at generation but also at structured analysis, comparison, and rubric-based evaluation [26]. Work on LLM-as-a-judge [66] shows that, when guided by explicit criteria and grounded in relevant evidence, their assessments align closely with human and expert judgements, and [17] applies this property to sign-to-text translation quality. In parallel, tool-augmented agentic frameworks extend these capabilities, allowing models to query specialised modules, compare heterogeneous outputs, retrieve external knowledge, and reason across multiple evidence sources rather than relying on a single monolithic forward pass [48, 61]. Within SL specifically, SignAgent [9] applies agents to data annotation and curation. SL evaluation is a natural fit for this agentic paradigm, since no single representation or model captures all relevant aspects of quality: grammaticality, phonological accuracy, motion fluency, and visual fidelity each demand different forms of evidence and analysis. Rather than treating the LLM as the recogniser or generator itself, our framework uses it as a reasoning layer that interprets and cross-references the outputs of specialised tools together with structured linguistic knowledge. This yields a more transparent, multi-dimensional evaluation, closer to the way human raters assess signing in practice and aligned with the SLP tiers.

3 Method

We present BackTranslation2.0 (BT2), a deterministic agentic framework for linguistically grounded evaluation of text-to-sign translation. Rather than relying

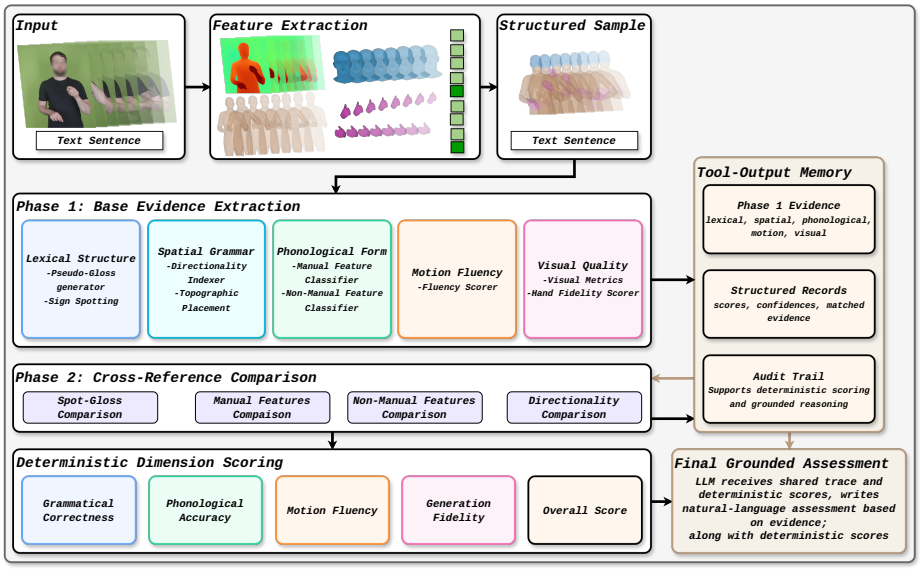

Fig. 1: Overview of the BackTranslation2.0 (BT2) pipeline. Given a source sentence s and generated sign output $\hat{y}$, multi-modal extraction produces a structured sample for segment-level and sequence-level analysis. Phase 1 base tools extract lexical, spatial, phonological, motion, and visual evidence, which is stored in a shared memory trace. Phase 2 comparison tools cross-reference this evidence against linguistic expectations to produce grounded judgements. Deterministic aggregation yields per-dimension scores and an overall understandability score, and a final LLM stage uses the same trace to write a grounded natural-language assessment.

on a single end-to-end metric, BT2 runs a fixed two-phase pipeline of specialised tools that assess *grammatical correctness*, *phonological accuracy*, *motion fluency*, and *generation fidelity*. It operates in two phases: Phase 1 extracts lexical, spatial, phonological, motion, and visual evidence from the generated signing. Phase 2 cross-references that evidence against linguistic expectations and retrieved lexical knowledge to produce grounded dimension-level judgements. All intermediate outputs are retained in a shared memory trace, enabling deterministic scoring and an auditable final natural-language assessment. Full implementation details for all tools are provided in the supplementary material.

3.1 Overview

Given a source sentence s and a generated sign output $\hat{y}$, BT2 returns per-tool structured evidence, per-dimension scores, a deterministic overall understandability score, and a grounded natural-language assessment. The generated sign output may be represented as RGB video, pose sequence, mesh animation, or a hybrid rendering format. BT2 first applies multi-modal feature extraction to

construct a shared structured sample,

$$\mathcal{X} = \Phi(s, \hat{y}) = \{x^{(1)}, \ldots, x^{(M)}, x^{\text{seq}}\}, \tag{1}$$

where $x^{(1)}, \ldots, x^{(M)}$ denote temporal segments of the signing stream and x^{seq} denotes sequence-level features over the full production. This allows some tools to operate at the *segment level* (*e.g.* sign-level lexical or phonological analysis) and others at the *sequence level* (*e.g.* fluency or global visual quality). BT2 then executes a fixed set of N=10 base tools, followed by a fixed set of comparison tools:

$$\mathcal{T}_{\text{base}} = \{T_1, \ldots, T_N\}, \qquad \mathcal{T}_{\text{comp}} = \{C_1, \ldots, C_L\}. \tag{2}$$

Unlike open-ended tool-using agents, BT2 does not decide dynamically which tools to call. Every sample follows the same deterministic pipeline, which makes the procedure reproducible, comparable across samples, and directly aligned with the target evaluation dimensions.

Tool-output memory. Each tool writes a structured record into a shared memory trace,

$$\mathcal{M} = \{m_1, \ldots, m_{N+L}\}, \tag{3}$$

where each memory entry stores a score, confidence, and evidence. The comparison stage consumes this memory rather than reprocessing the raw input, and the trace is passed to the final assessment stage. BT2 therefore provides an auditable path from low-level observations to final scores and explanation.

3.2 Phase 1: Base Evidence Extraction

Phase 1 extracts the evidence required for each quality dimension. The base tools target lexical content and ordering for grammar, manual and non-manual form for phonology, motion statistics for fluency, and visual plausibility for fidelity. In the primary extraction path, we extract SMPLX [35] for body pose with NLF [43], MANO [41] for hand pose with WiLoR [38], FLAME [27] for face parameters with TEASER [30], RTMPose [19] for whole-body tracking, Video-Depth-Anything [60] for depth, and $\text{SMPLX}_{\text{depth}}$ location features (SMPL-X re-optimised with depth cues), while segmentation defines temporal units [14]:

$$x^{(m)} = (b_m, h_m, f_m, \ell_m, r_m), \tag{4}$$

where $b_m, h_m, f_m, \ell_m, r_m$ denote segment-sliced body, hand, face, location, and whole-body features, and x^{seq} stores their full-sequence counterparts with segment set $\mathcal{S} = \{\tau_m\}_{m=1}^{M}$. Table 1 summarises the base tools, grouped below; full implementation and training details are in the supplementary material.

Grammar and spatial evidence. Pseudo-gloss generation maps the source sentence to candidate gloss-order sequences:

$$\mathcal{G} = f_{\text{pg}}(s), \quad g^{\star} = \text{First}(\mathcal{G}). \tag{5}$$

Table 1: Base tools used in Phase 1. 'Seg.' and 'Seq.' denote segment-level and sequence-level tools. The directionality tool uses a Sign Language Graph Convolutional Network (SL-GCN) [18] indexing classifier [40]. Use: Gram. = Grammatical Correctness, Phon. = Phonological Accuracy, Flu. = Motion Fluency, Fid. = Generation Fidelity.

Tool	Scope	Description	Use
Pseudo-Gloss	Seq.	LLM proposes BSL pseudo-gloss order.	Gram.
Sign Spotting	Seg.	Top-K sign-embedding gloss matches per segment.	Gram.
Directionality	Seg.	SL-GCN indexing/pointing classifier.	Gram.
Topographic Scorer	Seq.	FastDTW wrist-path direction matching.	Gram.
Handshape	Seg.	Transformer handshape classifier.	Phon.
Location	Seg.	Location classifier from $\text{SMPLX}_{\text{depth}}$.	Phon.
Non-Manuals	Seg.	AU-to-NMF classifier.	Phon.
Fluency	Seq.	Flow-matching bottleneck fluency score.	Flu.
Visual Metrics	Seq.	Deterministic smoothness/spectrum score.	Flu./Fid.
Hand Fidelity	Seq.	Binary hand-realism logit score.	Fid.

Here, f_{pg} is the pseudo-gloss generator, $\mathcal{G}$ is its candidate sequence set, and $g^\star$ is the top returned candidate used by downstream comparison tools. Sign spotting embeds each segment and returns top-K dictionary matches:

$$z_m = E_{\text{SR}}(\hat{y}_{\tau_m}), \quad \hat{g}_{m,1:K} = \text{TopK}_K \cos(z_m, E_{\text{dict}}) \,. \tag{6}$$

Here, $\hat{y}_{\tau_m}$ is the generated signing segment on interval τ_m, E_{SR} is the sign-embedding encoder [59], z_m is the segment embedding, E_{dict} is the lexical reference bank, K is the number of returned matches, and $\hat{g}_{m,1:K}$ are per-segment candidate glosses used by the Phase 2 spot-gloss comparison. For directionality, we combine SL-GCN [18] indexing prediction [40] with trajectory matching:

$$\hat{c}_m = \mathbf{1}\Big[f_{\text{SL-GCN}}(\tilde{b}_m, \tilde{h}_m) \geq \theta_{\text{idx}}\Big], \quad \hat{d} = \arg\min_{d \in \mathcal{R}} \text{DTW}(w_{xz}, r_d). \tag{7}$$

Here, $\tilde{b}_m, \tilde{h}_m$ are normalised body/hand features, θ_{idx} the indexing threshold, $\hat{c}_m$ the indexing decision for segment m, w_{xz} the wrist trajectory, and $\hat{d}$ the best-matching trajectory in the reference set $\mathcal{R}$ (giving r_d). The directional match rate is then mapped to the 0–4 topographic score used downstream.

Phonological evidence. Handshape and location use the same transformer architecture with spatial and temporal attention; the location branch adds contact-proximity features. The handshape classifier predicts both hands and keeps the dominant-confidence hypothesis:

$$p_m^L = f_{\text{hs}}(\nu(h_m^L)), \quad p_m^R = f_{\text{hs}}(\nu(h_m^R)), \quad (\hat{y}_m^{\text{hs}}, \hat{q}_m^{\text{hs}}) = \arg\max_{h \in \{L,R\}} \max p_m^h. \tag{8}$$

Here, h_m^L, h_m^R are left/right hand sequences, $\nu(\cdot)$ is hand-keypoint normalisation, f_{hs} is the handshape classifier, p_m^h is the class posterior for hand h, and $(\hat{y}_m^{\text{hs}}, \hat{q}_m^{\text{hs}})$ are the dominant handshape label and confidence. The location classifier uses $\text{SMPLX}_{\text{depth}}$ features with contact-proximity cues, applies two checkpoints of the same transformer backbone, and maps mean confidence to 0–4:

$$u_m = \pi_{\text{loc}}(\ell_m, \tau_m), \quad s_m^{\text{loc}} = 4 \cdot \frac{\max f_{\text{loc}}^a(u_m) + \max f_{\text{loc}}^b(u_m)}{2}. \tag{9}$$

Here, ℓ_m is the segment location tensor, $\pi_{\text{loc}}(\cdot)$ is the preprocessing map, u_m is the model input, $f^a_{\text{loc}}, f^b_{\text{loc}}$ are the two location classifiers, and s^{loc}_m is the 0–4 location score. The non-manual classifier maps per-frame action units (AUs) to NMF categories:

$$\mathcal{A}_m = \{a : \text{AU}_{t,a} > \eta_a,\ t \in \tau_m\}, \quad \hat{n}_m = \Gamma(\mathcal{A}_m; \mathcal{M}_{\text{AU}\to\text{NMF}}). \tag{10}$$

Here, η_a is the AU threshold for unit a, $\mathcal{A}_m$ is the active-AU set on segment τ_m, $\mathcal{M}_{\text{AU}\to\text{NMF}}$ is the semantic mapping, $\Gamma(\cdot)$ is the segment-level aggregation map, and $\hat{n}_m$ is the inferred non-manual label set.

Motion and fidelity evidence. For motion fluency, a conditional flow-matching model [31] encodes each modality sequence, and we use the U-Net bottleneck tensor for scoring. The bottleneck features are averaged over time and evaluated under a reference Gaussian in bottleneck space:

$$\begin{aligned} v_i &= \frac{1}{T_b}\sum_{t=1}^{T_b} B_{i,:,t}, \\ \log p(v_i) &= -\frac{1}{2}\left[(v_i-\mu)^\top \Sigma^{-1}(v_i-\mu) + \log|\Sigma| + C\log(2\pi)\right]. \end{aligned} \tag{11}$$

Here, $B_{i,:,t}$ is the U-Net bottleneck feature of sample element i at bottleneck time index t, T_b is bottleneck length, v_i is the mean bottleneck vector, (μ, Σ) are the reference Gaussian parameters in C dimensions, and $\log p(v_i)$ is the fluency evidence prior to 0–4 calibration. The log-probability is linearly calibrated to $[0, 4]$, and hand/body branches are averaged when available. Visual metrics provide deterministic kinematic evidence from body pose:

$$c_{\text{vis}} = w_{\text{hf}}\phi_{\text{hf}} + w_{\text{sm}}\phi_{\text{sm}} + w_{\text{act}}\phi_{\text{act}} + w_{\text{fr}}\phi_{\text{fr}}, \tag{12}$$

where $\phi_{\text{hf}}, \phi_{\text{sm}}, \phi_{\text{act}}, \phi_{\text{fr}}$ encode high-frequency energy, smoothness (jerk), motion activity, and dominant-frequency plausibility, $w_\cdot$ are fixed weights, and c_{vis} is the aggregated visual-motion evidence score. Hand fidelity trains a binary classifier on synthetic vs. real hand crops; the score is the mean logit over crops:

$$z_k = f_{\text{hfid}}(I_k), \quad s_{\text{hand}} = \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}} z_k. \tag{13}$$

Here, I_k is crop k, f_{hfid} is the hand-fidelity (real-vs-synthetic) classifier, z_k is the real-vs-synthetic logit for crop k, $\mathcal{K}$ is the crop set, $|\mathcal{K}|$ is the number of crops, and s_{hand} is the mean logit (linearly mapped to 0–4 for aggregation). Full architecture, training, and calibration settings are provided in the supplementary material.

3.3 Phase 2: Cross-Reference Comparison

Phase 2 maps Phase 1 outputs to grounded comparison evidence by contrasting expected and detected signals under tool-specific context. Instead of averaging independent predictions, this stage enforces linguistically structured cross-referencing before deterministic scoring. Phase 2 uses four fixed comparison tools:

spot-gloss comparison, manual-features comparison, non-manual-features comparison, and directionality comparison.

$$\mathcal{E}_j = \mathcal{C}_j(a_j, b_j \mid \kappa_j), \quad j \in \{\text{sg}, \text{man}, \text{nmf}, \text{dir}\}. \tag{14}$$

Here, sg, man, nmf, and dir denote spot-gloss, manual-features, non-manual-features, and directionality comparison tools, respectively; a_j denotes expected-side evidence, b_j denotes detected-side evidence from Phase 1, κ_j denotes tool-specific context (dictionary resources, rules, and prompt constraints), and $\mathcal{E}_j$ denotes comparison evidence written to memory.

Tool instantiations. The four comparison tools instantiate this general form as:

$$\begin{aligned}
\mathcal{E}_{\text{sg}} &= \mathcal{C}_{\text{sg}}\big(g^\star, \{\hat{g}_{m,1:K}\}_{m=1}^{M} \mid \kappa_{\text{sg}}\big)\,, \\
\mathcal{E}_{\text{man}} &= \mathcal{C}_{\text{man}}\big((g^\star, \mathcal{E}_{\text{sg}}), \{(\hat{y}_m^{\text{hs}}, s_m^{\text{loc}})\}_{m=1}^{M} \mid \kappa_{\text{man}}\big)\,, \\
\mathcal{E}_{\text{nmf}} &= \mathcal{C}_{\text{nmf}}\big((s, g^\star), \{\hat{n}_m\}_{m=1}^{M} \mid \kappa_{\text{nmf}}\big)\,, \\
\mathcal{E}_{\text{dir}} &= \mathcal{C}_{\text{dir}}\Big((s, g^\star), (\{\hat{c}_m\}_{m=1}^{M}, \hat{d}) \mid \kappa_{\text{dir}}\Big)\,.
\end{aligned} \tag{15}$$

Here, $\mathcal{E}_{\text{sg}}$ is spot-gloss semantic alignment evidence, $\mathcal{E}_{\text{man}}$ is manual phonological compatibility evidence, $\mathcal{E}_{\text{nmf}}$ is non-manual expectation-vs-detection evidence, and $\mathcal{E}_{\text{dir}}$ is indexing/topographic directionality evidence. Concretely, $\mathcal{C}_{\text{sg}}$ uses deterministic prefiltering followed by LLM semantic matching on unresolved pairs, $\mathcal{C}_{\text{man}}$ is dictionary-based with optional LLM fallback for low-confidence mapping, $\mathcal{C}_{\text{nmf}}$ can infer expected markers via LLM or deterministic rules, and $\mathcal{C}_{\text{dir}}$ is deterministic. The resulting evidence set $\mathcal{M}_{\text{comp}} = \{\mathcal{E}_{\text{sg}}, \mathcal{E}_{\text{man}}, \mathcal{E}_{\text{nmf}}, \mathcal{E}_{\text{dir}}\}$ is consumed by deterministic dimension scoring.

3.4 Deterministic Dimension Scoring

After both phases, BT2 deterministically computes four dimension scores, $\mathbf{d} = [d_{\text{gram}}, d_{\text{phon}}, d_{\text{flu}}, d_{\text{fid}}]$, each on a common 0–4 scale. Each dimension is formed from fixed subcomponent weights, with renormalisation when a subcomponent is unavailable. No LLM step modifies these numeric scores.

Overall understandability. The authoritative overall score $u = \frac{1}{4}(d_{\text{gram}} + d_{\text{phon}} + d_{\text{flu}} + d_{\text{fid}})$ is the arithmetic mean of the four deterministic dimensions.

3.5 Final Grounded Assessment

The final stage receives the full memory trace $\mathcal{M}$ together with fixed scores $\mathbf{d}$ and u. Its role is explanation only: it produces a grounded natural-language assessment and does not recompute or replace any numeric score. BT2 therefore returns structured per-tool evidence, fixed per-dimension scores, a fixed overall score, and an auditable grounded assessment.

4 Experiments

We evaluate BT2 on a purpose-built BSL benchmark (Sec. 4.1) probing the categories our metric targets: grammatical correctness, phonological accuracy, motion fluency, generation fidelity, and overall understandability. We compare BT2 against reference-free and reference-based baselines using human ratings of understandability and overall quality (Sec. 4.2), and test whether it is sensitive to the same controlled anomalies and corruptions identified by human raters. We then ablate each component (Sec. 4.3) and demonstrate large-scale generalisation on a synthetic corpus of procedurally corrupted sentences (Sec. 4.4).

4.1 Evaluation Dataset

Our benchmark has four complementary subsets spanning real signing, controlled corruption, synthetic production, and targeted spatial-grammar evaluation, summarised in Table 2.

Human Sentence Repetition Tasks (SRT). This subset contains real BSL sentence-repetition recordings from six signers at different proficiency levels (three learner levels and native), each producing the same six sentences. It provides paired comparison against native reference signing and captures natural proficiency-driven quality variation.

Known Corruptions. To isolate specific linguistic failure modes, we record human signers producing targeted corruptions of the same six sentences: missing non-manual features, incorrect word order, and incorrect handshape. This subset tests whether metrics penalise specific anomalies rather than only broad visual changes.

Synthetic Productions. We evaluate a set of synthetic productions for the same reference content, spanning text-to-sign and re-animation systems across pose-based, mesh-based, and RGB-based production systems, including SL-specific approaches SignGAN [46], SignSplat [16], SignStream [52], and SignSparK [31], as well as human motion generation systems, GUAVA [64] and Wan2.1 [58]. This subset focuses on production-method quality under shared content.

Directionality and topographic placement. We include a targeted set of sentences probing spatial grammar, including directional verbs, pronominal indexing, and topographic placement in signing space. This subset stresses phenomena that are weakly captured by conventional automatic metrics. Together, these subsets cover natural proficiency variation, controlled linguistic corruption, synthetic production quality, and spatial-grammatical correctness.

Human study. We conduct a user study with 20 participants (deaf and hearing) on a benchmark subset. Participants provide 5-point Likert ratings for *understandability* and *overall quality*, and flag perceived anomalies (handshape, hand positioning, non-manual features, word order, visual fidelity, fluency, and jitter). This supports both human-alignment analysis and corruption-sensitivity analysis. The full protocol, developed in collaboration between linguists and deaf experts, is provided in the supplementary material.

Table 2: Composition of the evaluation benchmark. Further details are provided in the supplementary material.

Subset	Content	Purpose
Human SRT	6 signers × 6 sentences	Proficiency variation
Known Corruptions	6 sent. × 6 sign. × 3 corrupt.	Anomaly sensitivity
Synthetic Productions	6 production systems	Production-method evaluation
Directionality / Topography	Targeted spatial-grammar set	Grammatical placement

4.2 Results

Table 3 compares BT2 with reference-free and reference-based baselines using direction-normalised Pearson r and Spearman ρ agreement with human labels. Pearson r captures linear agreement in score magnitude, while Spearman ρ captures rank-order agreement. Direction normalisation is applied so higher values always indicate stronger alignment with human judgements, including for distance-style baselines. The same test cases can therefore show different r and ρ values when magnitude spacing and rank ordering disagree. Reference-free methods rely on source-sentence or back-translated text signals, while reference-based methods compare against a target signing realisation in motion or latent space. BT2 combines category-level linguistic and visual evidence into deterministic scores, and uses the final language model to provide explanation text.

Against reference-free baselines, BT2 Overall is strongest in three of the four columns: *Known Corruptions* and both synthetic columns. *Human SRT* is the only column where BT2 Overall trails reference-free baselines. On *Known Corruptions*, BT2 Overall reaches '1.00 / 1.00' and matches or exceeds the strongest reference-based alternatives.

The per-component block shows that BT2 remains competitive with reference-based methods in key settings. BT2 Phonology reaches '0.87 / 0.90' on *Human SRT*, matching the best reference-based Spearman and clearly exceeding the strongest reference-free result. In this real SRT subset, sentence content is fixed and capture conditions are matched, so raters can separate signer proficiency primarily through manual and non-manual form differences, which aligns with the strong phonology result. BT2 Fidelity is strongest on synthetic quality ('0.83 / 0.94') and remains above both reference-free baselines and the strongest reference-based alternative on that synthetic-quality column. By contrast, BT2 Overall ('-0.43 / -0.30') and BT2 Fidelity ('-0.81 / -0.90') are weaker in *Human SRT*, suggesting that global-quality judgements in this subset are influenced by factors beyond signer-linguistic skill. BT2 Grammar is also weaker in *Human SRT* ('-0.45 / -0.60') because natural sign production allows more flexible word placement than tightly controlled corruption settings.

Human SRT is intrinsically difficult for rank-based evaluation because of score compression and low inter-rater agreement: raters exhibit a ceiling effect (74% of ratings $\geq$4), signer means span only a 0.55-point window on the 1–5 scale, and ICC$\approx$0. BT2's phonological signal (ρ=0.90) succeeds where genuine variation exists, while the negative Overall most likely reflects the difficulty

Table 3: Alignment between automatic metrics and human judgements across known corruptions, synthetics, and human SRT. 'Under.' and 'Qual.' denote understandability and overall quality. Metric cells show direction-normalised Pearson r / Spearman ρ (higher is better). **Bold** marks column-wise best values, and green marks BT2 cells that match or exceed the strongest reference-free baseline. DTW-MPJPE is the Mean Per-Joint Position Error after Dynamic Time Warping alignment, and SignCLIP is evaluated in pose-to-pose (p2p) and pose-to-text (p2t) configurations. The lower blocks report BT2 per-component scores and human inter-rater agreement (Gwet's AC2 [13], ICC [51], within-1, Quad. weighted κ [8], and Krippendorff's α [24] for anomaly flags).

Method	Known Corrupt.	Synthetics		Human SRT
	Under.	Under.	Qual.	Under.
Reference-based (req. source sequence)				
DTW-MPJPE	0.91 / 0.50	-0.14 / -0.03	-0.33 / -0.37	0.04 / 0.00
SiBLEU [67]	-0.58 / -0.50	0.03 / 0.31	-0.05 / 0.31	**0.96** / **0.90**
SignCLIP [21] p2p				
BSL sentence	-0.97 / -1.00	**0.85** / 0.90	0.82 / 0.70	0.69 / 0.80
BSL segmentation	-0.55 / -0.50	0.75 / 0.70	0.63 / 0.60	-0.46 / -0.40
Multilingual sentence	0.85 / **1.00**	0.70 / 0.70	0.55 / 0.60	-0.10 / -0.40
Multilingual segmentation	-0.59 / -0.50	0.83 / 0.70	0.75 / 0.60	-0.30 / 0.00
SVAE [10]	0.95 / **1.00**	0.43 / 0.26	0.54 / 0.31	0.81 / 0.80
Reference-free (req. source text)				
BT1 [45]				
BLEU-4 [34]	-0.19 / 0.00	-0.38 / -0.44	-0.15 / -0.10	0.00 / 0.00
BLEURT [49]	0.09 / 0.50	-0.35 / 0.03	-0.39 / -0.09	0.15 / 0.35
ChrF [37]	-0.87 / -1.00	-0.27 / -0.09	-0.54 / -0.37	-0.15 / -0.35
ROUGE [29]	-0.19 / 0.00	-0.34 / -0.44	-0.08 / -0.10	0.00 / 0.00
SignCLIP [21] p2t				
BSL sentence	0.25 / 0.50	0.03 / 0.30	0.17 / 0.40	0.56 / 0.60
BSL segmentation	-0.25 / -0.50	-0.33 / -0.30	-0.20 / -0.10	-0.78 / -0.70
Multilingual sentence	0.89 / 0.50	-0.64 / -0.60	-0.41 / -0.30	-0.07 / -0.10
Multilingual segmentation	0.77 / 0.50	-0.37 / -0.10	-0.50 / -0.30	-0.11 / -0.30
BT2 Overall (ours)	**1.00** / **1.00**	0.42 / 0.60	0.52 / 0.66	-0.43 / -0.30
Per-component				
BT2 Overall	**1.00** / **1.00**	0.42 / 0.60	0.52 / 0.66	-0.43 / -0.30
BT2 Phonology	0.28 / 0.50	0.48 / 0.26	0.35 / -0.03	0.87 / **0.90**
BT2 Grammar	-1.00 / -1.00	-0.50 / -0.14	-0.45 / -0.26	-0.45 / -0.60
BT2 Fluency	0.28 / 0.50	0.21 / 0.37	0.40 / 0.54	-0.46 / -0.40
BT2 Fidelity	0.80 / **1.00**	0.74 / **0.94**	**0.83** / **0.94**	-0.81 / -0.90
Human inter-rater agreement				
Gwet's AC2	0.69	0.69	0.68	0.69
ICC(2,1)	-0.02	0.51	0.35	0.07
Within-1 agreement	0.74	0.79	0.78	0.75
Quadratic weighted κ	-0.03	0.46	0.35	0.09
Handshape flag α	0.24	0.26	0.26	0.09
Non-manual flag α	0.27	0.29	0.29	-0.03
Word-order flag α	0.30	0.01	0.01	-0.01

Table 4: BT2 component ablation versus human ratings on Known Corruptions (KC), Synthetics (SC), and Human SRT. Cells report direction-normalised r / ρ; red (↓) marks a drop and green (↑) an improvement over the full system. †On Human SRT all ablations remain negative, indicating a structural rather than component-specific effect. The Grammar sub-score's $\rho = -1.00$ on KC reflects over-penalised flexible BSL word order, corrected to $+1.00$ by full aggregation.

Condition	KC	Synthetics (SC)		SRT†
	Under. r/ρ	Under. r/ρ	Qual. r/ρ	Under. r/ρ
No LLM (word heuristic)	.73(↓27%)/.50(↓50%)	.21(↓50%)/.37(↓38%)	.43(↓17%)/.54(↓18%)	−.68(↓58%)/ − .40(↓33%)
Phase 1 only	.73(↓27%)/.50(↓50%)	.21(↓50%)/.37(↓38%)	.43(↓17%)/.54(↓18%)	−.68(↓58%)/ − .40(↓33%)
Phase 1 + det. Phase 2	.08(↓92%)/.50(↓50%)	.22(↓48%)/.09(↓85%)	.42(↓19%)/.37(↓44%)	−.63(↓47%)/ − .50(↓67%)
Full − spot-gloss	.73(↓27%)/.50(↓50%)	.31(↓26%)/.37(↓38%)	.41(↓21%)/.49(↓26%)	−.68(↓58%)/ − .40(↓33%)
Full − manual feats.	.56(↓44%)/.50(↓50%)	.26(↓38%)/.37(↓38%)	.34(↓35%)/.49(↓26%)	−.66(↓53%)/ − .70(↓133%)
Full − non-manual	.35(↓65%)/.50(↓50%)	.50(↑19%)/.54(↓10%)	.54(↑4%)/.60(↓9%)	−.09(↑79%)/ − .10(↑67%)
Full − topo. dir.	.88(↓12%)/.50(↓50%)	.46(↑10%)/.43(↓28%)	.52(↑0%)/.60(↓9%)	−.34(↑21%)/ − .10(↑67%)
BT2 (ours)	**1.00/1.00**	**.42/.60**	**.52/.66**	**−.43/ − .30**

of reliable ranking in this highly compressed regime. Notably, *Known Corruptions*—also human-signed, but with a signer *deliberately* introducing handshape, location, and word-order errors—reaches ρ=1.00, confirming that BT2 responds to signing quality rather than signing origin.

The inter-rater block clarifies why grammatical alignment is harder to capture. Word-order flag agreement is highest in *Known Corruptions* ($\alpha = 0.30$), but near zero in *Synthetics* ($\alpha = 0.01$) and *Human SRT* ($\alpha = -0.01$), where word-order variation is less overt and consistently labelled. This pattern is consistent with greater word-placement flexibility in natural signing, and helps explain why grammar-focused correlations are weaker than phonological ones.

Figure 2 measures anomaly alignment directly. BT2 Final shows the strongest score ('0.66'), ahead of SignCLIP p2p ('0.32') and SiBLEU ('0.27'), and Panel B shows the expected negative trend between BT2 score and anomaly rate.

4.3 Ablation

We ablate seven variants of BT2 against human ratings to isolate the contribution of each component and of the language model itself (Table 4). Each variant either removes a base or comparison tool or replaces the Phase 2 language model with a deterministic fallback; full definitions are provided in the supplementary material. Every ablation reduces the Known-Corruptions Spearman ρ from 1.00. Replacing the Phase 2 language model with deterministic fallbacks reduces synthetic-understandability ρ by 85% (from 0.60 to 0.09), showing the gain comes from reasoning, not engineering alone. Removing it entirely (No LLM) leaves only Phase 1, since both calls are essential to comparison. Crucially, the negative *Human SRT* Overall correlation persists across all seven ablations, with no variant winning on any category; this confirms the negative result is structural to the compressed-rating subset rather than a removable component. The per-component sub-scores (lower block of Table 3) are complementary rather than redundant: Fidelity reaches ρ=0.94 on *Synthetics*, while Phonology reaches ρ=0.90 on *Human SRT*, where proficiency shows through articulatory form.

4.4 Corruption Sensitivity and Generalisation

Beyond the human-rated benchmark, we test corruption sensitivity at scale on a separate synthetic corpus: we generate BSL Corpus [47] sentences from continuous segments and corrupt them across five modality types (handshape, location, word order, temporal, and fidelity), retaining 900 matched pairs (n=180) scoring $\geq$2.5/4 under BT2; details are in the supplementary material. Table 5 reports win rates conditioned on which classifier fires. Across all five corruption types, BT2 ranks the uncorrupted sequence above its corruption on the overall score (65–79%) and on every matched target dimension: handshape (79–88%), location (73–98%), and grammar (56–61%). Fluency is deliberately wrong-way for the handshape and fidelity corruptions, which alter phonological form without disrupting motion, confirming BT2 localises each degradation to its target dimension. Where no difference is detected, BT2 abstains rather than penalising.

Table 5: Per-dimension win rates on synthetic BSL Corpus corruptions (uncorrupted BT2 ≥2.5, n=180), conditioned on which classifier fires. **Dark** ≥80%; mid 65–79%; light 55–64%; *red* wrong-way. **Bordered** cells mark the overall score and matched target dimensions; columns are the measured categories (overall, phonology, handshape/location accuracy, grammar, fluency, fidelity) and rows the corruption type *(left)*.

Corr.	n	Ovrl	Phon.	HS	Loc	Gram	Flu.	Fid.
Handshape classifier fires:								
Handshape	138	76%	73%	**86%**	75%	60%	*28%*	65%
Location	129	75%	76%	**82%**	74%	61%	52%	59%
Word-order	134	75%	73%	**82%**	73%	58%	43%	58%
Temporal	121	79%	71%	79%	75%	60%	55%	66%
Fidelity	137	74%	73%	**85%**	73%	60%	*26%*	64%
Location classifier fires:								
Handshape	87	69%	68%	**87%**	**98%**	57%	*22%*	62%
Location	93	68%	72%	**83%**	**95%**	61%	51%	62%
Word-order	88	78%	73%	**88%**	**94%**	56%	47%	64%
Temporal	78	78%	67%	**87%**	**96%**	56%	51%	69%
Fidelity	88	65%	64%	**85%**	**97%**	59%	*27%*	66%

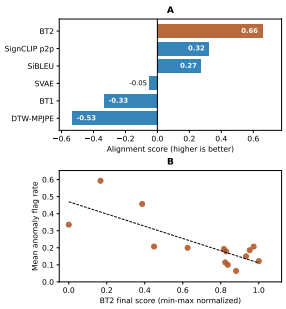

Fig. 2: Anomaly-alignment summary for the primary comparison set. Panel A: higher values indicate stronger penalisation of videos with more flagged anomalies. Panel B: BT2 final score against total anomaly rate across the 14 active questions *(right)*.

Table 6: Sentence-level directional predictions across clause pairs. **Bold** words are directional verbs; green pronouns are extracted grammatical-person entities; arrows ($1 \rightarrow 2$) denote predicted agreement direction between person loci (indices 1/2/3 = first/second/third person). NDV: no directional verb.

Sentence	Predicted	Matches GT
I will **help** you learn some signs but you can **ask** him	$1 \rightarrow 2$, $2 \rightarrow 3$	✓
I will **help** you learn some signs but I can **ask** him	$1 \rightarrow 2$, $1 \rightarrow 3$	✓
I will **help** him learn some signs but you can **ask** him .	$1 \rightarrow 3$, NDV	✓
She will **help** you learn some signs but he can **ask** me .	NDV, $1 \rightarrow 3$	✓
I will **help** you learn some signs but you can **ask** him	NDV, NDV	✓

4.5 Qualitative Analysis

Figure 3 illustrates per-tool behaviour: the Handshape and Contact-location columns clearly separate good from corrupted production, the Non-manuals column reflects the presence or absence of expected facial markers, and the hand-fidelity panel separates a reference clip from degraded synthesis, supporting the per-tool classification behind the quantitative results. Table 6 provides sentence-level directional-verb evidence for individual comparison tools. Across the shown clause pairs, the extracted person-locus sequences match the intended directional patterns, illustrating that the directional stack captures clause-conditioned role transfer rather than only surface motion similarity. Further qualitative material, including the motion-fluency visualisation and a full reasoning trace, is provided in the supplementary material.

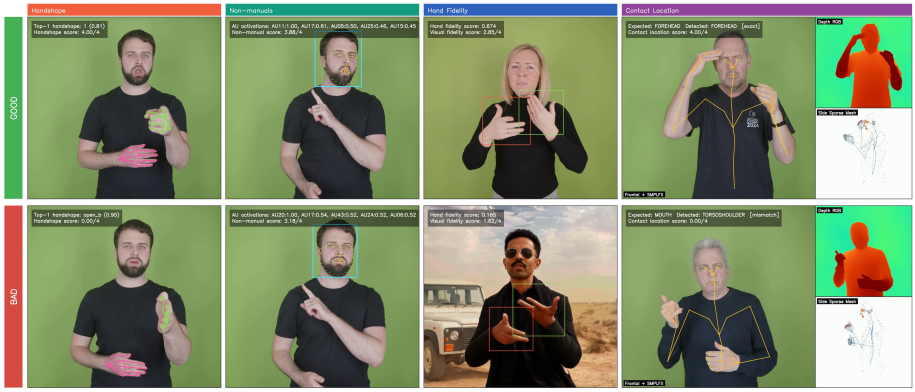

Fig. 3: Qualitative evaluation of BT2 visual-base tools; overlays show predicted class and score, with the top row rewarding good production and the bottom row penalising poor production. Columns (left to right): Handshape, Non-manuals, Hand fidelity, Contact location. Handshape and Non-manuals compare best production against matched corruptions; Hand fidelity contrasts a reference clip with poor-handed synthesis; Contact location shows correct (top) versus missing (bottom) contact via a frontal SMPL-X overlay, depth map, and side sparse view.

5 Conclusion

We introduced BackTranslation2.0 (BT2), a linguistically grounded metric for SLP that decomposes assessment into interpretable dimensions aligned with human rater criteria. It aligns more closely with human judgements than existing reference-free and reference-based baselines, gives substantially richer diagnostic insight, and captures phenomena conventional metrics miss—spatial grammar, directionality, and phonological mismatches. Beyond correlation gains, it provides interpretable evidence at multiple levels, from individual tool outputs to category-level scores, supporting not only benchmarking but also the analysis of failure modes in production systems. By grounding evaluation in structured visual and linguistic evidence rather than text surrogates or surface motion similarity, it offers a more faithful and practically useful framework for SLP assessment.

Acknowledgements

This work was supported by EPSRC grant APP24554 (SignGPT-EP/Z535370/1), EPSRC grant APP78083 (UMCS UKRI3927) and through funding from Google.org via the AI for Global Goals scheme. The authors acknowledge the use of Isambard-AI National AI Research Resource (AIRR) funded by UK DSIT via UKRI and STFC [ST/AIRR/I-A-I/1023]. This work reflects only the authors' views and the funders are not responsible for any use that may be made of the information it contains.

References

1. Adamo-Villani, N., Wilbur, R.B.: Asl-pro: American sign language animation with prosodic elements. In: Universal Access in Human-Computer Interaction. Access to Interaction. pp. 307–318. Springer International Publishing (2015)
2. Al-khazraji, S., Dingman, B., Lee, S., Huenerfauth, M.: At a different pace: Evaluating whether users prefer timing parameters in american sign language animations to differ from human signers' timing. Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility (2021)
3. Baltatzis, V., Potamias, R.A., Ververas, E., Sun, G., Deng, J., Zafeiriou, S.: 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). pp. 1985–1995 (2024)
4. Bowden, R., Saunders, B., Wheatley, M., Crowley, C., Hirshman, M., Birtles, D.: Taxonomy and Definitions for Terms Related to Automatic Translation of Spoken Language into Sign Language -Version 1.2 (2025)
5. Brentari, D.: A prosodic model of sign language phonology. Mit Press (1998)
6. Camgöz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural Sign Language Translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7784–7793 (2018)
7. Camgöz, N.C., Koller, O., Hadfield, S., Bowden, R.: 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 (CVPR). pp. 10023–10033 (2020)
8. Cohen, J.: Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin **70**(4), 213 (1968)
9. Cory, O., Sincan, O.M., Bowden, R.: SignAgent: Agentic LLMs for linguistically-grounded sign language annotation and dataset curation. arXiv preprint arXiv:2603.19059 (2026)
10. Cory, O., Sincan, O.M., Vowels, M., Battisti, A., Holzknecht, F., Tissi, K., Sidler-Miserez, S., Haug, T., Ebling, S., Bowden, R.: Modelling the Distribution of Human Motion for Sign Language Assessment. In: Proceedings of the European Conference on Computer Vision Workshops (ECCVW). pp. 1–19 (2025)
11. Crasborn, O., van der Kooij, E., Waters, D., Woll, B., Mesch, J., Bergman, B.: Frequency Distribution and Spreading Behavior of Different Types of Mouth Actions in Three Sign Languages. Sign Language & Linguistics **11**(1), 45–67 (2008)
12. Gong, J., Foo, L.G., He, Y., Rahmani, H., Liu, J.: LLMs are Good Sign Language Translators. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18362–18372 (2024)
13. Gwet, K.L.: Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology **61**(1), 29–48 (2008)
14. He, L.J., Walsh, H., Sincan, O.M., Bowden, R.: Hands-on: Segmenting individual signs from continuous sequences. In: 2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–5 (2025)
15. Huenerfauth, M., Marcus, M., Palmer, M.: Generating American Sign Language classifier predicates for English-to-ASL machine translation. Ph.D. thesis, University of Pennsylvania (2006)
16. Ivashechkin, M., Mendez, O., Bowden, R.: SignSplat: Rendering Sign Language via Gaussian Splatting. arXiv preprint arXiv:2505.02108 (2025)

17. Jang, Y., Raajesh, H., Momeni, L., Varol, G., Zisserman, A.: Lost in Translation, Found in Context: Sign Language Translation with Contextual Cues. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8742–8752 (2025)
18. Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., Fu, Y.: Skeleton aware multi-modal sign language recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 3413–3423 (2021)
19. Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., Chen, K.: RTMPose: Real-time multi-person pose estimation based on MMPose. arXiv preprint arXiv:2303.07399 (2023)
20. Jiang, Z., Leong, C., Moryossef, A., Cory, O., Ivashechkin, M., Tarigopula, N., Zhang, B., Göhring, A., Rios, A., Sennrich, R., Ebling, S.: Meaningful Pose-Based Sign Language Evaluation. In: Proceedings of the Tenth Conference on Machine Translation. pp. 64–80. Proceedings of the Association for Computational Linguistics (ACL) (Nov 2025)
21. Jiang, Z., Sant, G., Moryossef, A., Müller, M., Sennrich, R., Ebling, S.: SignCLIP: Connecting Text and Sign Language by Contrastive Learning. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 9171–9193 (Nov 2024)
22. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Transactions on Graphics **42**(4) (July 2023)
23. Kipp, M., Nguyen, Q., Heloir, A., Matthes, S.: Assessing the deaf user perspective on sign language avatars. In: The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility. pp. 107–114 (2011)
24. Krippendorff, K.: Computing Krippendorff's alpha-reliability. Communication Methods and Measures **5**(1), 1–12 (2011)
25. Kusters, A., De Meulder, M., O'Brien, D.: Innovations in deaf studies: Critically mapping the field. In: Innovations in deaf studies: The role of deaf scholars, vol. 12, pp. 1–53. Oxford University Press Oxford (2017)
26. Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., Liu, Y.: Llms-as-judges: a comprehensive survey on llm-based evaluation methods. arXiv preprint arXiv:2412.05579 (2024)
27. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM TOG **36**(6), 194:1–194:17 (2017)
28. Li, Z., Zhou, W., Zhao, W., Wu, K., Hu, H., Li, H.: Uni-Sign: Toward Unified Sign Language Understanding at Scale. In: Proceedings of the International Conference on Learning Representations (ICLR) (2025)
29. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics (ACL) (Jul 2004)
30. Liu, Y., Zhu, L., Lin, L., Zhu, Y., Zhang, A., Li, Y.: Teaser: Token enhanced spatial modeling for expressions reconstruction. In: ICLR (2025)
31. Low, J., Symeonidis-Herzig, A., Ivashechkin, M., Sincan, O.M., Bowden, R.: SignSparK: Efficient multilingual sign language production via sparse keyframe learning. arXiv preprint arXiv:2603.10446 (2026)
32. Napier, J., Leeson, L.: Sign language in action. In: Sign Language in Action, pp. 50–84. Palgrave Macmillan UK, London (2016)
33. Ong, S.C., Ranganath, S.: Automatic Sign Language Analysis: A Survey and the Future Beyond Lexical Meaning. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(6), 873–891 (2005)

34. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 311–318 (Jul 2002)
35. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: 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. pp. 10975–10985 (2019)
36. Pfau, R., Quer, J.: Nonmanuals: Their grammatical and prosodic roles, pp. 381–402. Cambridge University Press (2010)
37. Popović, M.: chrf: character n-gram F-score for automatic MT evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. pp. 392–395. Association for Computational Linguistics (Sep 2015)
38. Potamias, R.A., Zhang, J., Deng, J., Zafeiriou, S.: Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12242–12254 (2025)
39. Qi, F., Duan, Y., Zhang, H., Xu, C.: SignGen: End-to-End Sign Language Video Generation with Latent Diffusion. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 252–270 (2024)
40. Ranum, O., Hadfield, S., Bowden, R.: What's the point? spatial grammar & index resolution for sign language recognition. arXiv preprint arXiv:2606.08056 (2026)
41. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM TOG **36**(6), 245:1–245:17 (2017)
42. Sandler, W., Lillo-Martin, D.C.: Sign Language and Linguistic Universals. Cambridge University Press (2006)
43. Sárándi, I., Pons-Moll, G.: Neural localizer fields for continuous 3d human pose and shape estimation. Advances in Neural Information Processing Systems **37**, 140032–140065 (2024)
44. Saunders, B., Camgöz, N.C., Bowden, R.: Adversarial Training for Multi-Channel Sign Language Production. In: Proceedings of the British Machine Vision Conference (BMVC) (2020)
45. Saunders, B., Camgöz, N.C., Bowden, R.: Progressive Transformers for End-to-End Sign Language Production. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 687–705 (2020)
46. Saunders, B., Camgöz, N.C., Bowden, R.: Signing at Scale: Learning to Co-Articulate Signs for Large-Scale Photo-Realistic Sign Language Production. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5141–5151 (2022)
47. Schembri, A., Fenlon, J., Rentelis, R., Reynolds, S., Cormier, K.: Building the British Sign Language corpus. Language Documentation and Conservation **7**, 136–154 (2013)
48. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., Scialom, T.: Toolformer: Language Models Can Teach Themselves to Use Tools. In: NeurIPS. pp. 68539–68551 (2023)
49. Sellam, T., Das, D., Parikh, A.P.: BLEURT: Learning Robust Metrics for Text Generation. In: Proceedings 58th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 7881–7892 (2020)
50. Shen, X., Wang, X., Shen, L., Zhang, K., Yu, X.: Cross-View Isolated Sign Language Recognition via View Synthesis and Feature Disentanglement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)

51. Shrout, P.E., Fleiss, J.L.: Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin **86**(2), 420–428 (1979)
52. Signapse: Signstream api: Real-time british sign language translation. `https://www.signapse.ai/signstream-api` (2026), accessed: 2026-03-05
53. Starner, T., Weaver, J., Pentland, A.: Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **20**(12), 1371–1375 (1998)
54. Stoll, S., Camgöz, N.C., Hadfield, S., Bowden, R.: Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks. International Journal of Computer Vision **128**, 891–908 (2020)
55. Tang, S., He, J., Guo, D., Wei, Y., Li, F., Hong, R.: Sign-idd: Iconicity disentangled diffusion for sign language production. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 7266–7274 (2025)
56. Vogler, C., Metaxas, D.: Parallel Hidden Markov Models for American Sign Language Recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 116–122 (1999)
57. Walsh, H., Saunders, B., Bowden, R.: Sign stitching: A novel approach to sign language production. arXiv preprint arXiv:2405.07663 (2024)
58. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and Advanced Large-Scale Video Generative Models. arXiv preprint arXiv:2503.20314 (2025)
59. Wong, R., Camgoz, N.C., Bowden, R.: Signrep: Enhancing self-supervised sign representations. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22804–22814 (2025)
60. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth Anything V2. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) (2024)
61. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: ReAct: Synergizing Reasoning and Acting in Language Models. In: Proceedings of the International Conference on Learning Representations (ICLR) (2023)
62. Yin, A., Li, H., Shen, K., Tang, S., Zhuang, Y.: 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 (ACL). pp. 3345–3356 (2024)
63. Yin, A., Zhong, T., Tang, L., Jin, W., Jin, T., Zhao, Z.: Gloss Attention for Gloss-Free Sign Language Translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2551–2562 (2023)
64. Zhang, D., Liu, Y., Lin, L., Zhu, Y., Li, Y., Qin, M., Li, Y., Wang, H.: GUAVA: Generalizable Upper Body 3D Gaussian Avatar. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14205–14217 (October 2025)
65. Zhang, H., Shalev-Arkushin, R., Baltatzis, V., Gillis, C., Laput, G., Kushalnagar, R., Quandt, L.C., Findlater, L., Bedri, A., Lea, C.: Towards AI-driven Sign Language Generation with Non-manual Markers. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery (2025)
66. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems **36**, 46595–46623 (2023)

67. Zuo, R., Potamias, R., Sun, Q., Ververas, E., Deng, J., Zafeiriou, S.: MaDiS: Taming Masked Diffusion Language Models for Sign Language Generation. arXiv preprint arXiv:2601.19577 (01 2026)
68. Zuo, R., Potamias, R.A., Ververas, E., Deng, J., Zafeiriou, S.: Signs as Tokens: A Retrieval-Enhanced Multilingual Sign Language Generator. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 23806–23816 (2025)
69. Zuo, R., Wei, F., Mak, B.: Natural Language-Assisted Sign Language Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14890–14900 (2023)