DIRECT TRANSLATION BETWEEN SIGN LANGUAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Sign language translation has made substantial progress between sign and spoken languages, while translation across sign languages remains less explored. Translating directly between sign languages could support communication across signing communities without requiring a shared written language. A cascade of sign-to-text, spoken-language translation, and text-to-sign models offers one route, but can propagate intermediate errors and requires three sequential translation stages. We develop direct sign-to-sign translation, whose training is limited by the scarcity of parallel signing across languages. To address this obstacle, we adapt back-translation to construct cross-lingual pairs from existing text–sign corpora: the source signing is synthesized through a text bridge, while the target remains the gold sign from the original corpus. Using these pairs, we jointly train a single Qwen3-based model for text-to-sign and sign-to-sign translation. The latter generates target signing directly from source signing without an intermediate transcript. Experiments cover six directions among American, Chinese, and German Sign Language, using back-translation pairs and existing cross-lingual sign pairs. On synthetic sources, direct translation improves BLEU-4 in five directions and lowers overall motion error relative to the cascade. On existing test pairs, it improves BLEU-4 over the cascade by 1.91–3.87 points across all six directions.

1 INTRODUCTION

Deaf and hard-of-hearing signers from different countries face barriers to conversing in their native sign languages. Like spoken languages, sign languages have distinct grammars and vocabularies: American Sign Language (ASL), Chinese Sign Language (CSL), and German Sign Language (DGS) are different languages, and hundreds more are in active use worldwide (Yin et al., 2021). Such a conversation may rely on human interpreters, whose availability is limited, or on written text in a spoken language that the signers may not share. The latter route takes the conversation out of sign and can lose the spatial grammar, classifier predicates and prosody that signed dialogue relies on (Yin et al., 2021; De Coster et al., 2024). A system that translates directly between sign languages—taking one signer's clip in, returning an equivalent clip in another, with no detour through spoken-language text—could help address this barrier (Figure 1).

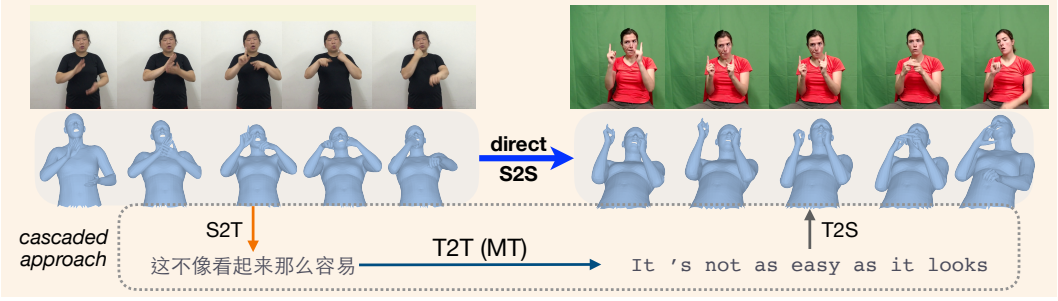

Figure 1: **Direct sign-to-sign translation.** Illustration of translation from CSL to ASL. The direct route (blue arrow) maps source signing to target signing without intermediate text. The cascaded route passes through source-language text, spoken-language MT, and target-language text before generating target signing.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The most natural way is a *cascade*: chain a sign-to-text (S2T) model (Camgöz et al., 2018; 2020; Lin et al., 2023; Yin & Read, 2020) with a spoken-language MT system (Liu et al., 2020; NLLB Team et al., 2022) and a text-to-sign (T2S) generator (Stoll et al., 2018; Saunders et al., 2020; 2021; Zelinka & Kanis, 2020; Zuo et al., 2025). This route reuses well-studied components, but it compounds errors across three stages, requires three sequential translation stages per query, and discards visual information not captured in the intermediate text.

A direct sign-to-sign (S2S) model bypasses this cascade, removing both the intermediate text bottleneck and the need to run three translation models in sequence (as in direct *speech-to-speech translation* (Jia et al., 2019; 2022; Rubenstein et al., 2023) in the spoken modality), but it is held back by the lack of parallel S2S data. The pioneering direct S2S work of Inan et al. (2025) automatically aligned three pairwise corpora (ASL $\leftrightarrow$ CSL, ASL $\leftrightarrow$ DGS, CSL $\leftrightarrow$ DGS) and trained a pose-to-pose Transformer based on Saunders et al. (2020). Their constructed data have two major limitations: *scale* and *alignment quality*. The training sets contain only 317–2,274 pairs per language pair. The authors themselves describe the construction as a "valuable yet noisy training set," acknowledging that some paired clips do not share the same meaning. In their limitations, they further explain that domain gaps between corpora make it impractical to identify equivalent clips from transcripts for longer, more complex meanings. Their reported BLEU-4, computed on text recovered from generated signing, is zero in two directions and at most 6.81 elsewhere.[1] Thus, direct S2S still needs a way to obtain substantially more paired supervision without relying on semantic overlap between independently collected corpora.

We address this data bottleneck by importing *back-translation* (BT) (Sennrich et al., 2016; Edunov et al., 2018)[2], a widely used technique for low-resource neural machine translation (NMT), into the sign modality: for each gold (text, sign) pair in a monolingual corpus, we translate the text into another spoken language, feed the result to a fixed T2S model to produce a synthetic source-sign clip, and pair that clip with the original gold sign as the target—yielding a large-scale parallel S2S training corpus (§2). For example, given an English–ASL pair $(\boldsymbol{x}^{\mathrm{en}}, \boldsymbol{y}^{\mathrm{ASL}})$ from How2Sign, we translate $\boldsymbol{x}^{\mathrm{en}}$ into Chinese $\hat{\boldsymbol{x}}^{\mathrm{zh}}$ via MT, and then use the T2S model to generate a synthetic CSL clip $\hat{\boldsymbol{y}}^{\mathrm{CSL}}$ from $\hat{\boldsymbol{x}}^{\mathrm{zh}}$; finally, we pair $(\hat{\boldsymbol{y}}^{\mathrm{CSL}}, \boldsymbol{y}^{\mathrm{ASL}})$ as a CSL $\rightarrow$ASL training instance. This construction does not require a matching source clip to exist in another corpus, and every target is a gold clip from the original corpus. Sign-side BT has previously been used in sign-to-text translation to augment sign–text training pairs (Zhou et al., 2021; Moryossef et al., 2021); here we use it to construct parallel training data across sign languages. Using our separately trained sign tokenizer, we train a single Qwen3-based model (Yang et al., 2025) jointly on T2S and S2S, sharing target-sign generation across natural text–sign pairs and our synthetic-source, gold-target pairs (§3).

Concretely, this paper makes three main contributions:

- We construct a large-scale parallel sign $\leftrightarrow$ sign training corpus of 112,324 ordered pairs by porting back-translation (Sennrich et al., 2016; Edunov et al., 2018) into the sign modality: a T2S model produces synthetic source clips from translated texts, which we pair with gold target clips from monolingual sign corpora (§2).
- While there exists a prior set of cross-lingual sign $\leftrightarrow$ sign dataset (Inan et al., 2025), it was aligned automatically and noisy by their own report; we extract a stricter diagnostic subset using an LLM judge, Sentence-BERT (Reimers & Gurevych, 2019), and manual transcript screening to reduce semantic mismatches in the paired references (§4).
- Direct translation achieves higher BLEU-4 than the cascade in all six Inan-Test directions (§5). On synthetic sources, it outperforms the cascade in five directions on both BLEU and motion error, except DGS$\rightarrow$ASL. It is $2.6\times$ faster than the cascade's summed stage times.

[1] These automatic scores also depend on the sign-to-text evaluator and tokenization, as discussed by Inan et al. (2025).

[2] Throughout this paper, *back-translation* refers to constructing synthetic-source training pairs, following Sennrich et al. (2016).

2 SYNTHETIC SIGN-TO-SIGN CORPUS VIA BACK-TRANSLATION

Large parallel sign ↔ sign corpora are scarce, so we use back-translation to construct synthetic-source, gold-target S2S pairs from text–sign corpora (Figure 2).

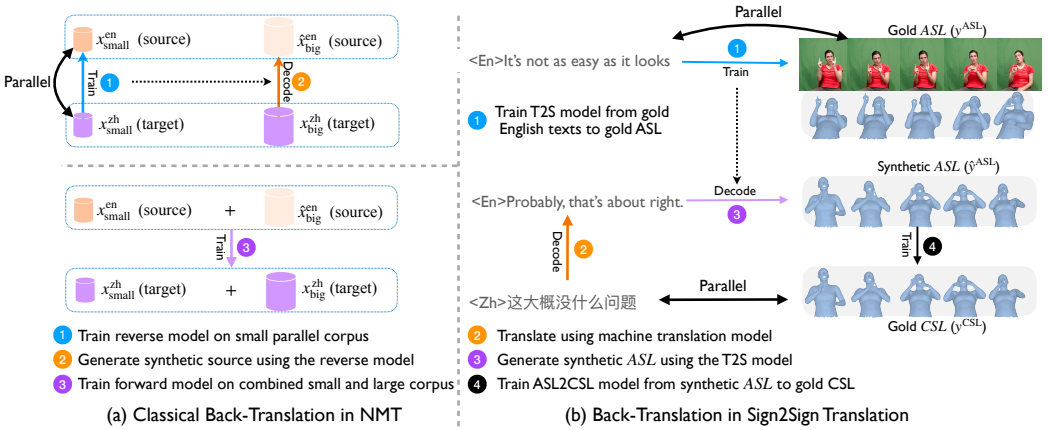

Figure 2: **Source-sign synthesis via a text bridge. (a)** Standard NMT synthesizes sources from gold target sentences (Sennrich et al., 2016). **(b)** We translate a gold CSL clip's transcript into English, generate an ASL source with T2S, and pair it with the original CSL target.

2.1 CLASSICAL BACK-TRANSLATION IN NMT

In NMT, back-translation expands scarce parallel data using abundant monolingual target-language text, without new human translations (Sennrich et al., 2016; Edunov et al., 2018).

Let $\boldsymbol{x}^{l}$ and $\boldsymbol{x}^{l'}$ denote target- and source-language sentences. A reverse model trained on the small parallel corpus maps each gold monolingual target $\boldsymbol{x}^{l}$ to a synthetic source $\hat{\boldsymbol{x}}^{l'}$. The pairs $(\hat{\boldsymbol{x}}^{l'}, \boldsymbol{x}^{l})$ augment the original parallel data for the forward model (Figure 2(a)). For example, to train a German-to-English model, we generate synthetic German sentences from monolingual English text using an English-to-German model. Each generated sentence is paired with its original English target. We likewise synthesize source signing while retaining gold target clips (Figure 2(b)).

2.2 ADAPTING BACK-TRANSLATION TO SIGN-TO-SIGN TRANSLATION

For sign language s, let $\boldsymbol{y}^{s}$ denote a sign sequence paired with a transcript $\boldsymbol{x}^{l(s)}$ in its associated spoken language $l(s)$. We use How2Sign for English–ASL (Duarte et al., 2021), CSL-Daily for Chinese–CSL (Zhou et al., 2021), and Phoenix-2014T for German–DGS (Camgöz et al., 2018). These corpora provide paired transcripts and clips for each sign language, but do not directly provide parallel sign pairs across languages.

For each gold text–sign pair $(\boldsymbol{x}^{l(s)}, \boldsymbol{y}^{s})$ and each desired source sign language $s' \neq s$, we translate the target transcript into $l(s')$ and synthesize source signing in s' (Figure 2b):

$$\hat{\boldsymbol{x}}^{l(s')} = \mathrm{MT}^{l(s)\to l(s')}(\boldsymbol{x}^{l(s)}), \qquad \hat{\boldsymbol{y}}^{s'} = \mathrm{T2S}^{s'}(\hat{\boldsymbol{x}}^{l(s')}). \tag{1}$$

We use TRANSLATEGEMMA 4B (Finkelstein et al., 2026) for MT; $\mathrm{T2S}^{s'}$ uses a fixed generator conditioned on sign language s'. The generator outputs SMPL-X motion sequences (Pavlakos et al., 2019) rather than raw video. Appendix A.1 gives the model and decoding settings. Each pair $(\hat{\boldsymbol{y}}^{s'}, \boldsymbol{y}^{s})$ has a synthetic source and gold target.

Together, MT and T2S take the role of the reverse model in classical back-translation. This construction requires neither a pre-existing S2S model nor frame-level sign alignment.

Applying this procedure to the official training splits yields 112,324 ordered pairs across the six directions among $\{\mathrm{ASL}, \mathrm{CSL}, \mathrm{DGS}\}$ (Table 1), with two source languages per target corpus. Test pairs are constructed separately (§4).

Table 1: **BT-constructed S2S training corpus (112,324 pairs).** Sources (s') are synthesized from translated text; targets (s) are gold clips from the corresponding corpus. Each count applies to both directions sharing a target corpus. Lengths are motion timesteps (four frames, three part codes each).

Direction ($s' \to s$)	Gold target corpus	Pairs/direction	src_len	tgt_len
ASL → CSL DGS → CSL	CSL-Daily	18,399	32.37 13.30	29.27
CSL → ASL DGS → ASL	How2Sign	30,671	27.28 15.03	36.63
ASL → DGS CSL → DGS	Phoenix-2014T	7,092	54.43 32.51	28.29

Earlier sign-side back-translation constructed sign–text pairs within one sign/spoken-language pair (Zhou et al., 2021; Moryossef et al., 2021); our construction pairs different sign languages. Keeping the target clip gold avoids synthesizing the supervision target, but MT and T2S errors can still affect the synthetic source, which may also omit signing content absent from the transcript. We therefore evaluate back-translated sources and existing source clips separately (§5).

3 Joint Text-to-Sign and Sign-to-Sign Modeling

We train one Qwen3-0.6B causal language model (Yang et al., 2025) on natural text–sign pairs and the sign–sign pairs constructed in §2. Both tasks generate target signing, conditioned on text for T2S or source signing for S2S. A separately trained, frozen VQ-VAE (van den Oord et al., 2017) represents signing as discrete motion tokens, giving the two tasks a shared output representation. The language model predicts target tokens from the corresponding source input, and the VQ decoder converts these predictions back to motion (Figure 3).

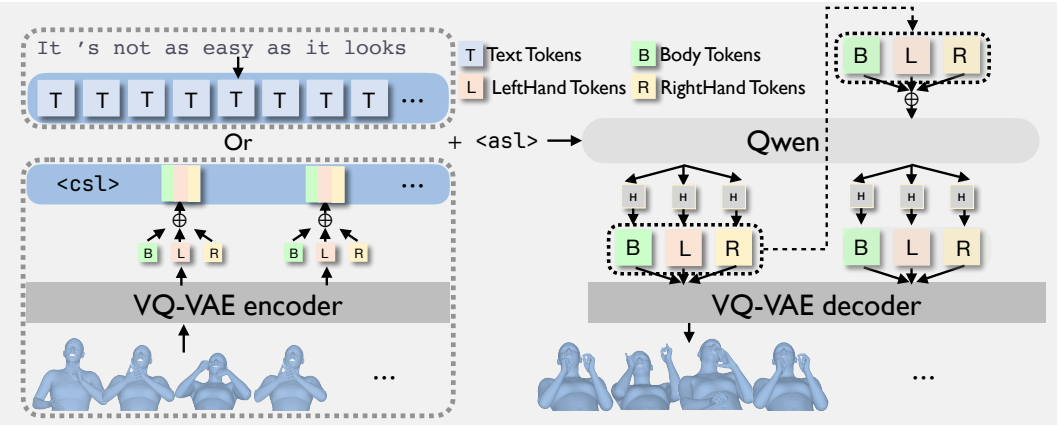

Figure 3: **Shared causal model for T2S and S2S.** Target motion is generated autoregressively without intermediate text in the S2S path. Training supplies preceding gold codes.

3.1 Shared Motion Tokens and Task-Specific Prefixes

Each frame's 133 SMPL-X parameters (Pavlakos et al., 2019) form a 43-channel body stream (30 body-pose, 3 jaw, and 10 expression) and two 45-channel hand streams. The shared VQ-VAE encodes the three streams with separate body, left-hand, and right-hand codebooks. Let $\boldsymbol{x}^{l(s)}$ be a transcript paired with signing in language s. After tokenization, we use $\boldsymbol{y}^s$ for the resulting motion-token sequence and $|\boldsymbol{y}^s|$ for its length in motion timesteps, including the end-of-sequence (EOS) token. A gold sign sequence is $\boldsymbol{y}^s = (\boldsymbol{y}^s_1, \ldots, \boldsymbol{y}^s_{|\boldsymbol{y}^s|})$, and $\hat{\boldsymbol{y}}^s$ denotes a synthetic sequence. At motion timestep t, three scalar token IDs form $\boldsymbol{y}^s_t = (y^{s,\text{body}}_t, y^{s,\text{left}}_t, y^{s,\text{right}}_t)$; a non-EOS timestep represents four

video frames. We model the two translation tasks with conditional probabilities

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}^s \mid \boldsymbol{x}^{l(s)}) \quad \text{and} \quad p_{\boldsymbol{\theta}}(\boldsymbol{y}^s \mid \hat{\boldsymbol{y}}^{s'}). \tag{2}$$

Here, s' and s denote distinct source and target sign languages, and $\boldsymbol{\theta}$ denotes the trainable language-model parameters. We use natural text–sign pairs for T2S and back-translated pairs (§2) for S2S. Both use gold targets; only the S2S training source is synthetic. At inference, S2S can use existing source clips.

We extend the Qwen vocabulary with motion-code tokens and the <asl>, <csl>, and <dgs> tags. At timestep t, the model's embedding table maps each token ID to a vector: the body embedding is $\boldsymbol{e}_t^{s,\text{body}}$, the left-hand embedding is $\boldsymbol{e}_t^{s,\text{left}}$, and the right-hand embedding is $\boldsymbol{e}_t^{s,\text{right}}$. We combine them into one input embedding $\boldsymbol{e}_t^s$ using scalar hand weight α:

$$\boldsymbol{e}_t^s = (1-2\alpha)\boldsymbol{e}_t^{s,\text{body}} + \alpha\boldsymbol{e}_t^{s,\text{left}} + \alpha\boldsymbol{e}_t^{s,\text{right}}. \tag{3}$$

We fix $\alpha = 0.4$, giving weights of 0.2 for body and 0.4 for each hand. In Figure 3, $\oplus$ denotes this weighted fusion, while $+$ denotes prefix concatenation. The sequence advances once per motion timestep; text uses the Qwen tokenizer and embeddings.

The T2S prefix is source text followed by the target sign-language tag. The S2S prefix contains the source tag, fused source motion, an end-of-source token, and the target tag (Figure 3). Each generated target position attends to the complete source prefix and preceding target positions.

3.2 Joint Training Objective

Both tasks train the same target-token predictor. With teacher forcing, the model conditions each prediction on the source prefix and preceding gold target tokens. Following SOKE's simultaneous part-wise prediction strategy (Zuo et al., 2025), we predict body, left-hand, and right-hand tokens at each motion timestep. Our implementation uses a shared vocabulary projection followed by three part-specific vocabulary masks and separate softmax normalization. Each H in Figure 3 denotes one such part predictor; the projection parameters are shared. Let $\boldsymbol{c}$ be the source prefix (text $\boldsymbol{x}^{l(s)}$ for T2S or synthetic signing $\hat{\boldsymbol{y}}^{s'}$ for S2S, with language tags) and $\boldsymbol{y}^s$ the gold target sequence. The conditional probability of $\boldsymbol{y}_t^s$ is the product of its body, left-hand, and right-hand token probabilities, all conditioned on the same prefix and the preceding target timesteps, denoted by $\boldsymbol{y}_{<t}^s$. The sequence loss is

$$\ell(\boldsymbol{c}, \boldsymbol{y}^s) = -\sum_{t=1}^{|\boldsymbol{y}^s|} \log p_{\boldsymbol{\theta}}(\boldsymbol{y}_t^s \mid \boldsymbol{c}, \boldsymbol{y}_{<t}^s), \tag{4}$$

This loss sums the three parts' cross-entropy losses over time. For each task, we sum $\ell(\boldsymbol{c}, \boldsymbol{y}^s)$ over the minibatch and divide by the total number of valid target timesteps, yielding $\mathcal{L}_{\text{T2S}}$ and $\mathcal{L}_{\text{S2S}}$. Source-prefix, language-tag, and padding positions are excluded. The tasks have equal weight: $\mathcal{L} = \mathcal{L}_{\text{T2S}} + \mathcal{L}_{\text{S2S}}$. Section 5.3 compares joint and single-task training.

3.3 Autoregressive Motion Decoding

At inference, the target tag initializes all three streams. At each timestep, the model selects body, left-hand, and right-hand codes, fuses their embeddings, and appends the result as input for the next prediction. EOS is masked in both hand streams; when the body emits EOS, both hands are forced to emit EOS at that timestep. The frozen VQ decoder converts output codes to motion; decoding settings are given in Appendix A.1.

4 Existing Sign-to-Sign Data Set and Our High-Quality Subset

Our evaluation uses two different types of datasets: (a) existing clips aligned across languages (see below) and (b) synthetic sources constructed from official held-out text–sign pairs (§2 and §5).

4.1 Existing Sign-to-Sign Cross-Lingual Pairs (Inan-Test)

Inan et al. (2025) align clips across ASL, CSL, and DGS by translating transcripts and matching them with a multilingual paraphrase model. Their released test set, Inan-Test, contains 1,781 pairs and 3,562

Table 2: Inan-Test and Inan-Strict: pair counts and transcript agreement. Strict ratings (1–5) and multilingual MiniLM-L12-v2 cosine means include repeated pair rows. Higher ratings indicate stronger event agreement.

	#Pairs		Qwen rating	S-BERT cos.
Language pair	Test requested	Strict	Strict	Strict
ASL ↔ CSL	677	190	4.95	0.88
ASL ↔ DGS	435	6	4.00	0.65
CSL ↔ DGS	669	213	4.19	0.67
All	1,781	409	4.54	0.76

directed requests before availability checks (Table 2). Similar transcripts need not convey equivalent signing, so mismatched references can penalize correct translations. We therefore complement this test set with a separate transcript-screened diagnostic, Inan-Strict.

4.2 Extracting High-Quality Cross-Lingual Subset (Inan-Strict)

To construct Inan-Strict, we pool the prior work's released train, development, and test pairs. For each transcript pair, Qwen3-8B (Yang et al., 2025) rates agreement on events, actions, and key entities with an integer score from 1 to 5 and provides a brief justification. We also compute cosine similarity between the transcripts' multilingual MiniLM embeddings (Reimers & Gurevych, 2019). We screen for ratings of at least 4 and similarity above 0.5, then retain selected exceptions after author discussion (Appendix A.2). This screens transcript agreement, not full linguistic equivalence of the signs.

Before availability checks, the 409 retained rows correspond to 394 distinct pairs and 224 source-clip–target-language inputs; an input can have multiple references. Only six pairs link ASL and DGS. Higher agreement scores reflect selection, not independent validation of alignment.

Because Strict draws from all three partitions, it is not a subset of Inan-Test. In 389 of its 409 rows, both clips occur in the underlying corpora's training annotations. Strict therefore serves as a diagnostic with training-clip overlap, not an unseen-clip test (Appendix A.2).

Figure 4 illustrates transcript pairs retained in the Strict subset and removed pairs for comparison.

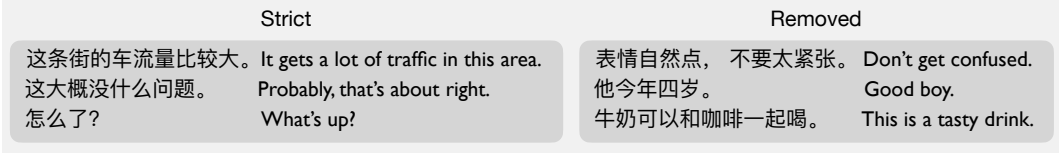

Figure 4: **Transcript pairs for CSL–ASL clips in the Strict subset.** Removed pairs are also shown for comparison.

5 Experiments

We compare direct S2S with an S2T→MT→T2S cascade across six directions among ASL, CSL, and DGS. Inputs include (a) existing clips from Inan-Test and Inan-Strict and (b) signs back-translated from held-out text–sign pairs (*BT-input*; §2). The cascade's S2T model depends on the input format:

- **Inan-Test and Inan-Strict:** Uni-Sign (Li et al., 2025) for ASL, Geo-Sign (Fish & Bowden, 2025) for CSL, and MMSLT (Kim et al., 2025) for DGS. The first two use pose inputs and the third uses RGB from the existing clips.
- **Back-translated source signs:** Generated SMPL-X motion is incompatible with these pose or RGB inputs. Preliminary training of alternative S2T models on our representation largely underperformed their published results and barely improved on our Sign Language Transformer (SLT) (Camgöz et al., 2020) baseline. We therefore use SLT, which accepts this motion, using the same model weights as the SLT evaluator for each source language.

Both cascades use TranslateGemma 4B for MT and direct's joint Qwen3-0.6B checkpoint and frozen sign tokenizer for T2S (Appendix A.1).

5.1 Evaluation Protocol and Inan-Test Comparison

Direct and cascade share clips, references, and target-language evaluators, though source representations can differ. Inan-Test has 3,542 common directed requests with paired targets. Strict has 219 source-clip–target-language anchors, with one prediction per system per anchor, scored against potentially multiple references (availability exclusions: Appendix A.2).

Metrics We prioritize pose-space error: tuned pose-distance metrics show stronger correlations with human judgments than S2T-based BLEU (Jiang et al., 2025). Following Zuo et al. (2025), we use mean per-joint position error (MPJPE) with Procrustes alignment and FastDTW (Salvador & Chan, 2007), denoted DTW-PA-MPJPE$_\downarrow$. We sum joint-averaged errors in metres along the DTW path without path-length normalization (Appendix A.4). On Strict, each system selects the reference with the lowest joint-weighted All DTW-PA-MPJPE and uses it for all body parts. Overall means weight inputs equally.

For comparison with prior work (Inan et al., 2025), we report BLEU-4$^\uparrow$ (Papineni et al., 2002) on text recovered by fixed target-language SLT evaluators, distinct from cascade front ends. Corpus BLEU uses all eligible reference transcripts on Strict and paired references on Inan-Test, without averaging directions. Gold-sign scores (Table 3) characterize evaluators on their own reference sets; they are neither upper bounds nor a cross-paper normalization.

Table 3: Direct and cascade results on Inan-Test (3,542 paired requests). MPJPE denotes cumulative DTW-PA-MPJPE in metres; B4 denotes BLEU-4. Inan's scores come from Inan et al. (2025), Tables 5–6, with different evaluators and evaluation sets. Gold-sign evaluator BLEU-4 (Inan / ours, each on its own reference set): ASL 2.93 / 12.11; CSL 2.25 / 15.62; DGS 12.36 / 16.59. Bold compares our paired systems.

Target	ASL				CSL				DGS			
Source	CSL		DGS		ASL		DGS		ASL		CSL	
Method	MPJPE$_\downarrow$	B4$^\uparrow$	MPJPE$_\downarrow$	B4$^\uparrow$	MPJPE$_\downarrow$	B4$^\uparrow$	MPJPE$_\downarrow$	B4$^\uparrow$	MPJPE$_\downarrow$	B4$^\uparrow$	MPJPE$_\downarrow$	B4$^\uparrow$
Inan	–	0.96	–	0.69	–	0.00	–	0.00	–	5.73	–	5.21
Cascade	**3.58**	7.57	3.27	4.25	3.70	8.10	3.95	6.12	1.95	6.29	2.47	9.68
Direct	3.89	**9.48**	**3.23**	**7.66**	**3.34**	**10.09**	**3.77**	**8.14**	**1.81**	**8.23**	**2.23**	**13.55**

Inan-Test comparison. With a shared evaluator, direct improves BLEU by 1.91–3.87 points over cascade in all six directions (Table 3). Its DTW-PA-MPJPE is lower overall (3.11 versus 3.22) and in five directions. CSL→ASL is the exception, with higher errors for every part; DGS→ASL also has higher hand errors despite lower All DTW-PA-MPJPE. Evaluator differences on gold signs prevent attributing the gap from Inan et al. (2025) solely to translation quality.

5.2 Direct Translation and Cascades

We also compare direct and cascade on Inan-Strict and back-translated inputs (Tables 4 and 5). Different examples and references prevent isolating the effect of synthetic versus existing sources.

Inan-Strict results. On Strict (Table 4), direct has lower aggregate DTW-PA-MPJPE (2.25 versus 2.44), but is worse on ASL→CSL and DGS→ASL. Direct has higher BLEU in five directions, but is lower on DGS→ASL (4.52 versus 5.62).

The 76 distinct CSL source clips have 31 reference transcripts, yet cascade produces one front-end transcript and one motion sequence per target language. Repeated references therefore cannot fully explain these constant outputs. Excluding CSL sources leaves 143 anchors, where direct has higher DTW-PA-MPJPE (2.45 versus 2.25). We retain all 219 anchors but cannot explain the constant outputs, limiting this comparison.

Table 4: Direct and cascade results on Inan-Strict (219 anchors). Cascade S2T uses Uni-Sign (Li et al., 2025) (ASL), Geo-Sign (Fish & Bowden, 2025) (CSL), and MMSLT (Kim et al., 2025) (DGS). The cascade uses the same MT and T2S models as in Table 5. MPJPE denotes cumulative DTW-PA-MPJPE in metres; B4 denotes BLEU-4. Bold compares paired systems. Strict overlaps training clips and has constant CSL front-end outputs; see text.

Target	ASL				CSL				DGS			
Source	CSL		DGS		ASL		DGS		ASL		CSL	
Method	MPJPE$_\downarrow$	B4$^\uparrow$	MPJPE$_\downarrow$	B4$^\uparrow$	MPJPE$_\downarrow$	B4$^\uparrow$	MPJPE$_\downarrow$	B4$^\uparrow$	MPJPE$_\downarrow$	B4$^\uparrow$	MPJPE$_\downarrow$	B4$^\uparrow$
Cascade	3.71	8.21	**1.59**	**5.62**	**1.40**	8.73	3.21	7.50	1.57	4.87	2.36	8.13
Direct	**2.06**	**9.87**	2.14	4.52	2.03	**10.96**	**2.88**	**10.25**	**1.47**	**6.79**	**1.79**	**12.66**

Table 5: BT-input results (8,246 back-translated sources). Cascade uses SLT (Camgöz et al., 2020) for S2T and the same MT and T2S models as in Table 4. MPJPE denotes cumulative DTW-PA-MPJPE in metres; B4 denotes BLEU-4. Bold compares paired systems.

Target	ASL				CSL				DGS			
Source	CSL		DGS		ASL		DGS		ASL		CSL	
Method	MPJPE$_\downarrow$	B4$^\uparrow$	MPJPE$_\downarrow$	B4$^\uparrow$	MPJPE$_\downarrow$	B4$^\uparrow$	MPJPE$_\downarrow$	B4$^\uparrow$	MPJPE$_\downarrow$	B4$^\uparrow$	MPJPE$_\downarrow$	B4$^\uparrow$
Cascade	4.35	7.43	**3.92**	**8.57**	3.39	8.62	3.50	8.15	2.63	9.84	2.57	10.06
Direct	**3.75**	**9.65**	4.26	8.43	**3.34**	**10.81**	**3.35**	**10.19**	**2.40**	**11.62**	**2.40**	**12.28**

BT-input results. On back-translated inputs (Table 5), direct improves BLEU by 1.78–2.22 points and lowers All DTW-PA-MPJPE in five directions. Its overall DTW-PA-MPJPE is 3.57 versus 3.70 for the cascade. DGS→ASL is the exception on both metrics: BLEU is 8.43 versus 8.57, and DTW-PA-MPJPE is 4.26 versus 3.92. For DGS→ASL, we generate the DGS source from the English transcript via German and compare the output with the paired gold How2Sign ASL. Reported English–ASL misalignment (Tanzer et al., 2024) may create source–target mismatches, though it does not alone explain the gap.

Direct–cascade latency. On 75 common inputs at batch size one, direct averages 270 ms versus 698 ms for the cascade's summed stage means: a 2.6× speedup (61% lower time). Both routes are timed from prepared input representations to generated motion, excluding loading, I/O, SMPL-X extraction, rendering, and scoring (Appendix A.6).

5.3 Front-End Training Data and Joint-Task Training

S2T training data. We test whether the front-end training distribution affects cascade performance on real versus synthetic inputs. With the SLT architecture, MT, and Qwen T2S fixed, we compare real, synthetic, and equal-sized mixed training data. Synthetic training works best on synthetic inputs; mixed training works best on real inputs (Appendix A.3). This tests SLT's distribution sensitivity, not the alternative front ends used on Inan-Test and Inan-Strict.

Joint versus single-task training. We compare the selected joint and single-task models to examine how performance changes across T2S and S2S tasks. Joint training lowers synthetic-source S2S motion error for all three target languages relative to either single-task model, but increases T2S error relative to T2S-only training (Table 6). Unmatched task-specific exposure and compute prevent isolating the training objective's causal effect (Appendix A.3.2).

5.4 Qualitative Analysis

Before inspecting predictions, we select 15 examples across six directions by sample ID and source-token length. Each shows synthetic input, direct, cascade, and gold-reference videos with identical rendering (Appendix A.5). Samples 011, 013, and 014 produce identical direct videos from different

Table 6: Joint versus single-task training: cumulative DTW-PA-MPJPE ↓. S2S inputs are synthetic; T2S-only performs S2S zero-shot. Dashes mark unevaluated tasks. Training exposure is unmatched.

	S2S			T2S		
Training	ASL	CSL	DGS	ASL	CSL	DGS
T2S only	4.95	3.42	4.10	3.66	2.87	2.25
S2S only	4.60	3.62	2.75	–	–	–
Joint T2S +S2S	**4.00**	**3.34**	**2.40**	3.94	3.25	2.53

inputs; we retain these failures. The examples do not establish semantic fidelity or preservation of non-manual grammar and spatial reference.

6 Related Work

Sign-to-text and text-to-sign translation. S2T methods include neural translation (Camgöz et al., 2018), joint recognition–translation transformers (Camgöz et al., 2020), improved sequence decoders (Yin & Read, 2020), and gloss-free models (Lin et al., 2023). Uni-Sign (Li et al., 2025) uses pose pretraining and optional RGB fusion; Geo-Sign (Fish & Bowden, 2025) uses hyperbolic pose representations; and MMSLT (Kim et al., 2025) uses multimodal language-model supervision. T2S methods generate signing through GANs (Stoll et al., 2018), progressive transformers (Saunders et al., 2020), motion primitives (Saunders et al., 2021), gloss-mediated word-level synthesis (Zelinka & Kanis, 2020), or discrete tokens (Zuo et al., 2025). Combined with spoken-language MT (Liu et al., 2020; NLLB Team et al., 2022; Finkelstein et al., 2026), these methods form the S2T →MT→T2S cascade compared with our direct model.

Multilingual sign processing. MLSLT (Yin et al., 2022) supports multilingual S2T; JWSign (Gueuwou et al., 2023) supplies multilingual data; and WMT-SLT (Müller et al., 2023) benchmarks sign/spoken-language pairs. Jiang et al. (2023) translate into SignWriting. These efforts connect signing with text or writing, whereas our task translates between sign languages.

Cross-lingual sign-to-sign translation. Inan et al. (2025) align text-paired clips across ASL, CSL, and DGS to train a pose-to-pose model. We use their pairs for evaluation (§4) but construct training pairs through back-translation (§2). Earlier sign-side back-translation augments sign–text data within one sign/spoken-language pair (Zhou et al., 2021; Moryossef et al., 2021); ours pairs different sign languages. Direct translation also follows calls for communication tools without a shared spoken-language intermediary (Yin et al., 2021; De Coster et al., 2024).

7 Limitations

Source errors and corpus coverage. Gold targets cannot remove MT or T2S errors in synthetic sources. Although inference bypasses text, the text bridge used to construct training pairs may lose non-manual or spatial information. Language and domain co-vary across our corpora, limiting generalization. We will explore OpenASL (Shi et al., 2022) and YouTube-ASL (Uthus et al., 2023) to assess corpus and alignment effects.

Data and comparison conditions. Strict is small, uneven, and transcript-aligned rather than naturally parallel signing; 389 of its 409 rows overlap training clips. It does not test unseen-clip generalization. Constant CSL front-end outputs further limit its cascade comparison. Training controls use selected runs with unmatched exposure rather than multi-seed estimates.

Metrics and evidence. Cumulative DTW-PA-MPJPE depends on length, S2T-based scoring introduces evaluator error, and evaluator differences limit cross-paper BLEU comparisons. Without human evaluation, automatic metrics and examples cannot establish comprehensibility, naturalness, or linguistic fidelity.

Offline inference and deployment. The model observes the complete source clip. Inference timing excludes extraction and rendering and does not establish conversational latency or deployment readiness. Partnership with the Deaf-community partnership is necessary.

Ethics Statement

We plan to release code and non-restricted corpus and clip indices. Generated data and examples will be released only where permitted by the applicable source licenses and permissions; this release excludes original CSL-Daily samples without authorization. The present evaluation uses automatic metrics and qualitative examples; it does not establish the accessibility or suitability of generated signing for use by signing communities. Appendix A.7 describes the planned release and source access conditions.

AI Use Statement

We used OpenAI Codex to polish writing, draft manuscript sections, find related work, and assist with research planning, evaluation code, and result analysis. Model-assisted translation, synthetic data generation, and transcript screening are described in Sections 2 and 4. The authors reviewed the AI-assisted work and take responsibility for the paper.

References

Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *CVPR*, pp. 7784–7793, 2018.

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *CVPR*, pp. 10023–10033, 2020.

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

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giró-i-Nieto. How2Sign: A large-scale multimodal dataset for continuous American Sign Language. In *CVPR*, pp. 2735–2744, 2021.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *EMNLP*, pp. 489–500, 2018.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, Markus Freitag, and David Vilar. TranslateGemma technical report. *arXiv:2601.09012*, 2026.

Edward Fish and Richard Bowden. Geo-Sign: Hyperbolic contrastive regularisation for geometrically aware sign language translation. In *NeurIPS*, 2025.

Shester Gueuwou, Sophie Siake, Colin Leong, and Mathias Müller. JWSign: A highly multilingual corpus of Bible translations for more diversity in sign language processing. In *Findings of EMNLP*, pp. 9907–9927, 2023.

Mert Inan, Yang Zhong, Vidya Ganesh, and Malihe Alikhani. How to align multiple signed language corpora for better sign-to-sign translations? In *NAACL-HLT*, pp. 4003–4016, 2025.

Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. Direct speech-to-speech translation with a sequence-to-sequence model. In *Interspeech*, pp. 1123–1127, 2019.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In *ICML*, pp. 10120–10134, 2022.

Zifan Jiang, Amit Moryossef, Mathias Müller, and Sarah Ebling. Machine translation between spoken languages and signed languages represented in SignWriting. In *Findings of EACL*, pp. 1706–1724, 2023.

Zifan Jiang, Colin Leong, Amit Moryossef, Oliver Cory, Maksym Ivashechkin, Neha Tarigopula, Biao Zhang, Anne Göhring, Annette Rios, Rico Sennrich, and Sarah Ebling. Meaningful pose-based sign language evaluation. In *WMT*, pp. 64–80, 2025.

Jungeun Kim, Hyeongwoo Jeon, Jongseong Bae, and Ha Young Kim. Leveraging the power of MLLMs for gloss-free sign language translation. In *ICCV*, pp. 21048–21058, 2025.

Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. Uni-Sign: Toward unified sign language understanding at scale. In *ICLR*, 2025.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. Gloss-free end-to-end sign language translation. In *ACL*, pp. 12904–12916, 2023.

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

Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. Data augmentation for sign language gloss translation. In *AT4SSL*, pp. 1–11, 2021.

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. Findings of the second WMT shared task on sign language translation (WMT-SLT23). In *WMT*, pp. 68–94, 2023.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. *arXiv:2207.04672*, 2022.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *ACL*, pp. 311–318, 2002.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, pp. 10975–10985, 2019.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP-IJCNLP*, pp. 3982–3992, 2019.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Frank. AudioPaLM: A large language model that can speak and listen. *arXiv:2306.12925*, 2023.

Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Progressive transformers for end-to-end sign language production. In *ECCV*, pp. 687–705, 2020.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Mixed SIGNals: Sign language production via a mixture of motion primitives. In *ICCV*, pp. 1919–1929, 2021.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *ACL*, pp. 86–96, 2016.

Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. Open-domain sign language translation learned from online video. In *EMNLP*, 2022.

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

Garrett Tanzer, Maximus Shengelia, Ken Harrenstien, and David Uthus. Reconsidering sentence-level sign language translation. In *EMNLP*, 2024.

David Uthus, Garrett Tanzer, and Manfred Georg. YouTube-ASL: A large-scale, open-domain American Sign Language–English parallel corpus. In *NeurIPS*, 2023.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, pp. 6306–6315, 2017.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv:2505.09388*, 2025.

Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. MLSLT: Towards multilingual sign language translation. In *CVPR*, pp. 5109–5119, 2022.

Kayo Yin and Jesse Read. Better sign language translation with STMC-transformer. In *COLING*, pp. 5975–5989, 2020.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. Including signed languages in natural language processing. In *ACL-IJCNLP*, pp. 7347–7360, 2021.

Jan Zelinka and Jakub Kanis. Neural sign language synthesis: Words are our glosses. In *WACV*, pp. 3384–3392, 2020.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *CVPR*, pp. 1316–1325, 2021.

Ronglai Zuo, Rolandos Alexandros Potamias, Evangelos Ververas, Jiankang Deng, and Stefanos Zafeiriou. Signs as tokens: A retrieval-enhanced multilingual sign language generator. In *ICCV*, pp. 23806–23816, 2025.

A SUPPLEMENTARY MATERIAL

This appendix provides implementation settings, evaluation details, component controls, and qualitative examples.

A.1 TRAINING, DECODING, AND CASCADE CONFIGURATION

We train a shared sign tokenizer separately and freeze it during T2S/S2S training. We train the joint Qwen3-0.6B model for 100 epochs. Training uses two GPUs, batch size 16 per GPU, and eight-step gradient accumulation, giving a nominal effective batch size of 256. We use seed 1234, AdamW with learning rate 2×10^{-4}, decay rates $(\beta_1, \beta_2) = (0.9, 0.99)$ for the first and second moments, zero weight decay, cosine decay to 10^{-6}, and gradient clipping at 1.0.

Evaluation uses greedy decoding with a maximum of 100 motion-token timesteps, test batch size 16, and seed 1234. Motion scoring uses Procrustes alignment and FastDTW with radius 2. We report cumulative DTW-PA-MPJPE without path-length normalization. The length analysis in Appendix A.4 examines the resulting dependence on sequence length.

BT source generation and text translation. BT sources are generated by the fixed T2S-only Qwen3-0.6B model compared in §5.3, separate from the joint translation model. It uses greedy decoding, a limit of 100 motion-token timesteps, and earlier termination at EOS. The intervening text translation uses `google/translategemma-4b-it` with its official structured chat template and language codes `zh-CN`, `en`, and `de`. Decoding is deterministic (`do_sample=false`, `num_beams=1`), with `max_new_tokens=128`, batch size 1, and BF16 on one GPU.

Language-specific sign-to-text front ends. For ASL, we use the pose-only Uni-Sign model with its official How2Sign checkpoint. For CSL, we use Geo-Sign initialized from the released CSL-News pretraining checkpoint and fine-tuned on CSL-Daily, with validation-based checkpoint selection. Both use WholeBody-133 pose inputs. For DGS, we use RGB-based MMSLT on Phoenix-2014T. The chosen front end is shared by the two directions with the same source language; the downstream MT and target-sign generation stages complete the cascade. Their standalone corpus-test BLEU-4 scores are 14.50 on How2Sign, 23.53 on CSL-Daily, and 23.77 on Phoenix-2014T, respectively; these measure source recognition, not end-to-end S2S quality. For synthetic sources, the cascade instead uses SLT front ends with the corresponding language-specific evaluation weights. Their decoding uses beam sizes 5, 5, and 3 and maximum output lengths of 100, 50, and 30 tokens for ASL, CSL, and DGS, respectively.

A.2 EVALUATION SPLITS, EXCLUSIONS, AND REFERENCE SCORING

The 112,324 training pairs use the official training splits of How2Sign, CSL-Daily, and Phoenix-2014T; BT-input evaluation uses their official test splits. Pairing is at the utterance level, with synthetic sources and gold targets as described in §2.

Inan-Test requests 3,562 directed translations. Direct produces 3,542 outputs, with 20 requests lacking source or target motion assets. The cascade produces all 3,562 outputs, of which 3,552 have target motion for scoring. The paired Inan-Test tables compare both systems on the same 3,542 requests, scoring each against its paired target reference.

Strict independently pools the prior work's train, development, and test partitions. Of its 409 rows, 389 have both clips in monolingual training annotations, one in development, and 19 in test. It is therefore a diagnostic on existing clips, not an unseen-clip test.

Four rows with Qwen ratings below 4 and 18 with cosine similarity at or below 0.5 were retained after author discussion. The fixed manifest contains 394 distinct clip pairs; 15 occur twice. Repeated pair rows do not add generation trials or references. Before availability checks, the six directions contain 24, 6, 70, 66, 6, and 52 anchors, respectively, in the column order of Table 4.

Four three-frame ASL clips cannot be loaded by the fixed evaluation pipeline and are excluded as inputs and references, affecting ten manifest rows. We also exclude one ASL→DGS source absent from Uni-Sign's released inputs from both systems' comparison, retaining the same clip as a

DGS→ASL reference. This leaves 219 anchors and 767 source–reference comparisons; the 409-row inventory is preserved.

Multi-reference selection and sample-weighted aggregation. Each source anchor produces one prediction. Both systems use the same candidate references, deduplicated by target clip. For each candidate, we compute cumulative DTW-PA-MPJPE using a separate alignment path for each part. Let C_{body} denote the body error, C_{left} the left-hand error, and C_{right} the right-hand error. Weighting by 12 body joints and 21 joints per hand gives the aggregate C_{all}:

$$C_{\text{all}} = \frac{12C_{\text{body}} + 21C_{\text{left}} + 21C_{\text{right}}}{54}. \tag{5}$$

Each system selects the reference minimizing C_{all} and uses that same reference for all reported parts; ties are broken by target clip ID. No path-length normalization is applied. Overall motion Avg weights directions by their evaluated anchor counts.

BLEU-4 uses all eligible reference transcripts in a corpus calculation for each direction. Hypotheses are recovered from generated motion by the same target-language evaluator for both systems; the cascade's intermediate text is not used for scoring. No overall BLEU is reported. The ASL→DGS and DGS→ASL directions have only five and six anchors, respectively, and anchors sharing references need not be statistically independent.

Table 7 reports both systems on this common subset.

Table 7: Direct and cascade motion errors by body part on Inan-Strict. The CSL-source cascade front end produces constant text (see main text). All entries use cumulative DTW-PA-MPJPE ↓; the body and both hands use the same selected reference for each prediction. Avg gives each evaluated anchor equal weight.

Method	Direction	Anchors	Body	Left hand	Right hand	All
Direct	ASL→CSL	66	5.60	0.48	1.55	2.03
	ASL→DGS	5	3.97	0.69	0.81	1.47
	CSL→ASL	24	4.53	1.35	1.37	2.06
	CSL→DGS	52	4.69	0.76	1.16	1.79
	DGS→ASL	6	6.78	2.12	2.06	2.14
	DGS→CSL	66	7.68	1.11	1.91	2.88
	Avg	219	5.89	0.88	1.54	2.25
Cascade	ASL→CSL	66	3.86	0.30	1.10	1.40
	ASL→DGS	5	3.53	0.95	1.08	1.57
	CSL→ASL	24	7.84	2.39	2.68	3.71
	CSL→DGS	52	5.46	1.55	1.41	2.36
	DGS→ASL	6	3.42	1.05	1.08	1.59
	DGS→CSL	66	8.81	1.14	2.09	3.21
	Avg	219	6.15	1.11	1.64	2.44

A.3 Front-End Training, Joint-Model Results, and Tokenizer Errors

The following tables report auxiliary control experiments. System identities are given explicitly because these controls address different questions from the direct–cascade comparison on Inan-Test and Inan-Strict.

A.3.1 Real, Synthetic, and Mixed Front-End Training

To test sensitivity to the source-sign distribution, we keep the SLT front-end architecture, MT model, and Qwen-based T2S model fixed. We train the front end on real signs, T2S-generated signs, or an equal-sized 50/50 mixture (Table 8). This control does not compare the Uni-Sign, Geo-Sign, or MMSLT front ends used for Inan-Test and Inan-Strict.

On synthetic test inputs, training the front end on synthetic signs gives lower motion error and higher BLEU than training on real signs. On real test inputs, the mixed-training model has the lowest motion

Table 8: Real, synthetic, and mixed training data for the SLT front end, with Qwen-based T2S. These auxiliary reported aggregates cover six directions and are separate from the paired cohorts in the main tables. DTW-PA-MPJPE is lower-is-better; B4 is higher-is-better.

	Real test inputs		Synthetic test inputs	
S2T training inputs	DTW-PA-MPJPE	B4	DTW-PA-MPJPE	B4
Real	8.23	7.26	8.20	7.18
Synthetic	8.49	6.86	7.78	7.85
Equal-sized 50/50 mixture	8.14	7.37	7.92	7.61

error and highest BLEU among these three settings. The preferred training distribution therefore depends on the test inputs; these controls do not establish that source-distribution mismatch is absent.

A.3.2 Joint-Model Motion and BLEU by Task and Target Language

Table 9 reports the selected Qwen3-0.6B checkpoint trained jointly on T2S and S2S. Evaluation covers 4,126 held-out T2S text–sign inputs and 8,246 matched synthetic-source S2S inputs with gold target signing. Target-language and overall motion aggregates weight their contributing inputs equally. Each S2S target-language row pools the two source languages; the six directional results appear in Table 5. "All" is the evaluator's reported aggregate, rather than an unweighted mean of the three displayed parts. S2S BLEU is reported by direction, without averaging two corpus scores into a target-language score.

For the joint-versus-single-task comparison in Table 6, the T2S-only and joint runs use a cosine learning-rate floor of 10^{-6}, while S2S-only uses 10^{-7}. All three runs use a nominal effective batch size of 256, but their selected checkpoints are not matched for task-specific exposure or compute.

Table 9: Joint T2S/S2S model: cumulative DTW-PA-MPJPE $_{\downarrow}$ and BLEU-4$^{\uparrow}$, separated by task and target language. N counts the evaluation inputs. Target-pooled S2S BLEU is not reported; its per-direction scores appear in Table 5.

Method	Task	Target	Body	Left hand	Right hand	All	N	BLEU-4
Joint	T2S	ASL	9.02	2.42	2.56	3.94	2308	8.08
	T2S	CSL	8.24	1.54	2.10	3.25	1176	9.04
	T2S	DGS	6.35	1.34	1.54	2.53	642	10.52
	S2S	ASL	9.03	2.49	2.64	4.00	4610	–
	S2S	CSL	8.44	1.50	2.28	3.34	2352	–
	S2S	DGS	6.09	1.16	1.54	2.40	1284	–

Compared with the cascade on synthetic-source inputs (Table 5), Direct has higher errors for all parts in DGS→ASL; its right-hand error is also higher for ASL→CSL (2.29 versus 2.20) and marginally for CSL→DGS (a difference below 0.01).

A.3.3 Shared VQ-VAE Reconstruction Error

Reconstruction measures the tokenizer's ability to represent gold motion; it does not directly establish the semantic fidelity of generated signing.

Our tokenizer uses one shared VQ-VAE with separate body, left-hand, and right-hand codebooks, rather than three VQ-VAEs as in SOKE (Zuo et al., 2025). The parts have 43/45/45 input channels and 96/192/192 codebook entries, with 512-dimensional embeddings and four-fold temporal down-sampling. We use CSL-derived mean and standard deviation for normalization. The pooled token corpus uses all 96/192/192 codes; this distinct-code coverage does not imply uniform usage or full coverage within every language.

Tokenizer training uses one GPU, batch size 320, a 500-epoch schedule, and AdamW with learning rate 2×10^{-4}, moment decay rates $(\beta_1, \beta_2) = (0.9, 0.99)$, zero weight decay, and cosine decay to 10^{-6}. Table 10 reports the shared tokenizer's frame-aligned PA-MPJPE, without DTW.

Table 10: Shared VQ-VAE reconstruction errors by language and body part. Values are frame-aligned PA-MPJPE$_\downarrow$ in millimetres, without DTW. "Hand" averages the two hands; "All" weights the 14 body joints and 21 joints per hand equally.

	ASL			CSL			DGS		
Method	All	Body	Hand	All	Body	Hand	All	Body	Hand
Shared VQ-VAE	9.83	19.17	6.72	9.69	22.84	5.31	11.31	24.75	6.83

A.4 Reference-Length Dependence of Cumulative DTW-PA-MPJPE

For the Qwen3-0.6B T2S-only checkpoint, the saved evaluator inputs provide reference lengths after evaluator preprocessing: 40–400 frames, rounded to multiples of four. Cumulative DTW-PA-MPJPE depends on both local joint errors and alignment-path length. Dividing DTW-PA-MPJPE by reference length is a sensitivity analysis, not exact path normalization.

Table 11: Spearman rank correlation ρ between evaluator reference-frame length and cumulative DTW-PA-MPJPE, before and after division by reference length. Brackets show percentile 95% intervals from 3,000 bootstrap resamples of evaluation examples.

Language/dataset	N	DTW-PA-MPJPE ρ	Reference-normalized ρ
ASL/How2Sign	2,308	0.935 [0.924, 0.945]	-0.160 [-0.200, -0.120]
CSL/CSL-Daily	1,176	0.863 [0.844, 0.879]	-0.201 [-0.256, -0.141]
DGS/Phoenix-2014T	642	0.928 [0.914, 0.940]	-0.081 [-0.164, -0.003]

After reference-length normalization, all three correlations are small and negative. Thus, these saved outputs provide no evidence that the score divided by reference length increases with longer reference sequences. They do not show that Qwen S2S quality is length-invariant. Generated lengths and DTW path lengths were not saved, so exact path-normalized error is unavailable from these saved scores.

A.5 Fixed-Case Selection and Video Rendering

We select examples before inspecting predictions: the lexicographically first sample and the shortest and longest source-token sequences in each direction, with sample-ID tie breaking and deduplication. This produces 15 examples, each with synthetic input, direct prediction, cascade prediction, and gold reference videos (60 videos in total). The supplementary package includes 56 videos, withholding the four gold references from original CSL-Daily samples. Generated videos are retained for all 15 cases.

Videos share a camera and renderer, retain full sequences, and use each dataset's nominal frame rate without smoothing or temporal alignment. Contact sheets independently show first/middle/last frames; reference panels use the evaluator's 40–400-frame preprocessing, without synchronization across panels. SLT readouts are model predictions, not human annotations.

A.6 Stage Timings and Batch-Size Effects

Measurements of direct S2S and the cascade's T2S stage use an H200 NVL MIG 3g.71gb partition, FP32 without autocast, batch size one, greedy decoding capped at 100 timesteps, no KV cache, five warm-up batches, and three repeats. The cohort selects the first 16 canonical anchors per direction, or all available anchors when fewer exist, then applies the source-availability exclusion described in Appendix A.2. This leaves 75 anchors and 225 timing measurements, weighted equally.

Direct averages 270.293 ms: 2.772 ms source VQ encoding, 264.654 ms autoregressive generation, and 2.867 ms target decoding and motion assembly. It starts with device-resident pose and excludes CPU preprocessing and host-to-device transfer. Cascade T2S averages 380.508 ms from in-memory translated text, including prompt construction, tokenization, host-to-device transfer, generation, and target decoding. It excludes source recognition and MT. Both exclude model loading, external I/O, SMPL-X extraction, rendering, and scoring.

The S2T and MT measurements use the same 75 inputs and three repeats at batch size one, averaging 117.499 and 200.090 ms. Adding the T2S mean gives 698.097 ms, a $2.58\times$ speedup for direct and a 61.3% reduction in mean inference time. This total sums separately timed stages; it is not a single uninterrupted pipeline measurement. The comparison starts from each route's prepared input representation, including RGB for MMSLT, and ends at generated motion, excluding extraction and rendering.

Batch-size-one cascade outputs differ from the formal batched outputs on two of 75 anchors: one produces 156 rather than 188 frames; the other has different motion values at the same 28-frame length. The previously excluded direct anchor produced 32 versus 144 frames, but its exclusion was based on source-input availability. Timing therefore does not establish identical outputs across batch sizes.

A.7 Planned Artifacts and Dataset Release Restrictions

We plan to release code, model configurations, and non-restricted corpus/clip indices. Generated cross-lingual data, evaluation inputs and references, and qualitative examples will be released only where permitted by the applicable source terms. We have not obtained authorization to redistribute original CSL-Daily samples and exclude them from the planned public release.

Dataset access and redistribution terms. How2Sign is distributed under CC BY-NC 4.0 with attribution and a non-commercial restriction.[3] CSL-Daily access requires a research agreement; its terms prohibit redistributing the original dataset without prior written consent.[4] Original video clips will be referenced through indices and their source access channels. Releases derived from How2Sign and Phoenix-2014T will follow their respective licenses and access terms, including applicable attribution and non-commercial restrictions. Public download access alone is not treated as permission to redistribute source material.

[3] How2Sign dataset license

[4] CSL-Daily release agreement