ATLAS: TIMED LEXICAL ANCHORS FOR SIGN LANGUAGE PRODUCTION AT SCALE

Anonymous authors
Paper under double-blind review

ABSTRACT

Glosses, ordered transcriptions of signs as written words, considerably improve sign language production (SLP) when used as a condition, but require expert annotation that large signing corpora lack. Glosses derived from the text alone do not reproduce this gain, as they carry no information about when each sign occurs. We instead extract timed lexical anchors from the signing itself with a sign-identity model trained on isolated sign dictionaries, and combine them with the sentence as a hybrid condition: the anchors supply the timing of the signs they cover and the sentence supplies the remaining content. At inference, the anchors are predicted from the sentence alone. The scheme requires no gloss annotation of the training corpus and therefore extends to large unannotated datasets. We extract and release SMPL-X parameters for more than 1,700 hours of video from YouTube-SL-25, and train ATLAS, a continuous-latent flow-matching generator, on more than 600 hours of ASL signing, substantially more than used in prior work. The anchors improve back-translation at every data scale we test: an anchored model trained on a fraction of the data surpasses a sentence-only model trained on all of it, and relative to each system's ceiling, ATLAS reaches more than twice the back-translation score of the strongest baseline. Finally, we audit the SLP evaluation protocol used in the literature with content-free baselines, outputs that carry no signing by construction. Reference-based distance metrics favour low-amplitude motion, to the extent that such outputs can score higher than state-of-the-art models. We propose a protocol that reports the model's performance grounded in these baselines and measures content with a public back-translation model.

1 INTRODUCTION

Sign language production (SLP) is the task of turning a spoken-language sentence into the corresponding signs, usually represented as a sequence of body, hand and face poses Arkushin et al. (2023); Zuo et al. (2025; 2026); Low et al. (2026). Signing conveys meaning through several articulators at once, and what a sign means depends on how these channels are coordinated and timed. Therefore, a useful generator has to model the whole body, render the hands and face precisely, and keep them synchronised.

Recent state-of-the-art approaches, such as SOKE (Zuo et al., 2025) and M^{3}T (Symeonidis-Herzig et al., 2026), quantise decoupled body, hand and face streams into discrete tokens and decode them with a finetuned language model. Quantisation fits language modeling well, but signing is continuous and fluid, and a restricted codebook drops the fine articulation on which the meaning depends. Passing a ground-truth clip through SOKE's tokenizer, with no generation involved, already removes a big part of what a sign back-translator can read from it (§4.3), which bounds any generator built on those tokens. Following recent work on motion generation (Chen et al., 2023; Zhu et al., 2025), we instead encode motion into a continuous latent sequence with a VAE (Kingma & Welling, 2014), and generate these latents with a flow-matching transformer conditioned on the input sentence.

Beyond representing motion faithfully, a generator needs a condition that tells it which signs to produce and when, and the sentence alone is a poor one. Signs do not follow the word order of the sentence, and the mapping between the two is not consistent across signers or sentences, which makes the raw sentence a weak and noisy signal to learn from. Glosses, an ordered transcript of the signs, give the generator the structure the sentence lacks: in our experiments, conditioning on

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.

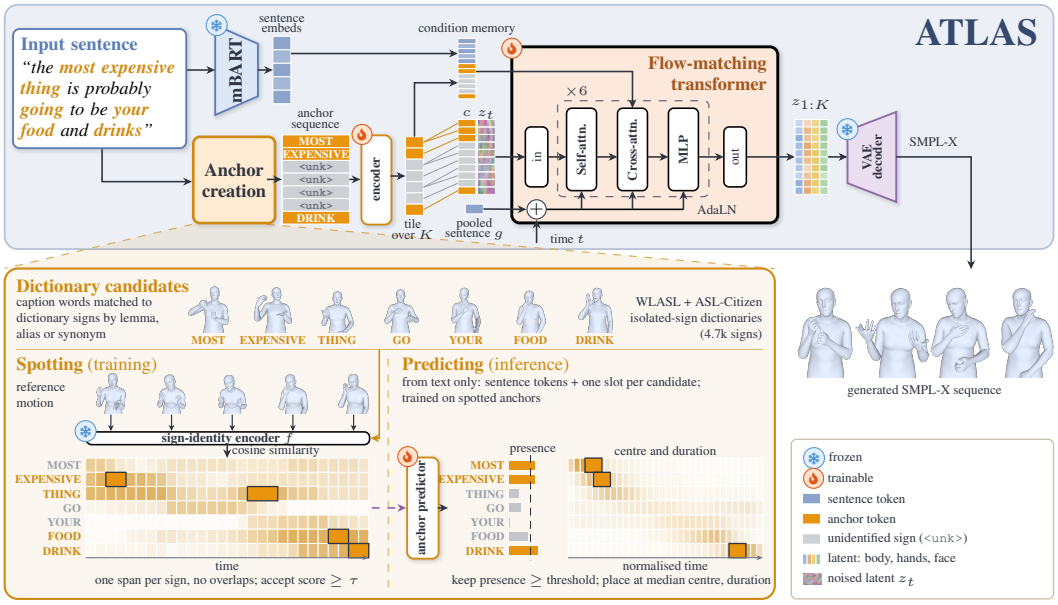


Figure 1: **ATLAS overview.** Anchor creation turns the sentence into a timed sequence of dictionary signs and <unk> slots. A flow-matching transformer reads these anchors together with the sentence embeddings and generates motion latents, which the VAE decodes to SMPL-X. During training, the anchors are spotted in the reference motion by a sign-identity encoder; at inference a predictor infers them from the text alone.

human-annotated glosses instead of the sentence more than halves the error rate of the generated signing (§4.1). However, glosses require expert annotation, and no large corpus has them. Deriving them from the text alone does not help either: caption keywords carry the sentence's order with no timing, and condition the generator no better than the sentence itself (§4.1). We therefore derive the lexical signal from the signing instead. A sign-identity model trained on isolated-sign dictionaries (Li et al., 2020; Desai et al., 2023) verifies which words of the caption are signed in the video and where, yielding *timed lexical anchors* without gloss annotation. Because a dictionary covers only a part of the signed vocabulary, the anchors are sparse, and the model consumes them through a *hybrid* condition in which the anchors supply the timing of the signs they cover and the sentence supplies the remaining content. Spotting requires the motion that is to be generated, so at inference a predictor infers the anchors directly from the sentence.

Since anchor extraction requires only sign motion, text captions and a sign dictionary, it extends to web-scale datasets without gloss annotation. We recover SMPL-X (Pavlakos et al., 2019) motion for more than 1,700 hours of YouTube-SL-25 (Tanzer & Zhang, 2024), with the face represented by FLAME parameters (Li et al., 2017) as in the SMPL-FX format of Symeonidis-Herzig et al. (2026), which we simply call SMPL-X. Dedicated per-frame estimators of the body, hands and face (Sárándi & Pons-Moll, 2024; Potamias et al., 2025; Liu et al., 2025) are combined into this model and refined against 2D whole-body keypoints (Yang et al., 2023) under temporal regularisation (§A), and we release the fitted parameters. Close to 600 hours of ASL signing from YouTube-SL-25 together with How2Sign provide substantially more ASL training data than prior SLP work. Back-translation of the generated signing, read by a public sign translator, rises with the amount of data, and the anchors improve it at every data size, to the extent that an anchored model trained on a small fraction of the data surpasses a sentence-only model trained on all of it. Relative to each system's ceiling, ATLAS reaches more than twice the back-translation score of the strongest baseline, and misplacing the anchors in time drops the score to the sentence-only level, confirming that timing is what they provide.

Measuring SLP performance reliably is itself difficult. The two standard metrics are a joint distance to the reference motion after dynamic time warping (DTW) and back-translation, the score of a sign translator that maps the generated motion back to text. Using *content-free baselines*, outputs that carry no signing by construction, we find that DTW favours low-amplitude motion near the

reference's mean pose. A frozen body can score better than published state-of-the-art models, while a correct rendition of the same sentence by another signer scores worse. Back-translation is closer to what SLP aims for, but most works train their own translator on small datasets and rarely report its score on the ground truth. Therefore, we propose a protocol that reports every score alongside content-free baselines and measures content with a public sign translation model (Li et al., 2025), read against its own ceiling on the ground truth.

In summary, our contributions are: (1) *timed lexical anchors*, spotted without gloss annotation by a dictionary-grounded sign-identity model calibrated against a wrong-caption control, and predicted from the sentence at inference; (2) *hybrid conditioning*, which combines sparse anchors with the sentence; (3) *SMPL-X parameters for YouTube-SL-25*, more than 1,700 hours of temporally optimised signing motion, which we release; and (4) *an audit of SLP evaluation* with content-free baselines, with a reporting protocol built on a shared public sign translator, read through a calibrated SMPL-X-to-keypoint adapter that we release.

2 Related Work

Sign language production. Early neural SLP was gloss-based: Progressive Transformers (Saunders et al., 2020) and Signing at Scale (Saunders et al., 2022) translate text to glosses and glosses to pose, the latter stitching dictionary signs, and both depend on gloss annotation for training and on a text-to-gloss translator at inference. Gloss-free systems followed: NSA (Baltatzis et al., 2024), T2S-GPT (Yin et al., 2024), S-MotionGPT (Hwang et al., 2024) and, most recently, SOKE (Zuo et al., 2025) and M^{3}T (Symeonidis-Herzig et al., 2026), which quantise decoupled body, hand and face streams into tokens and decode them autoregressively with a pretrained multilingual language model. SOKE also retrieves dictionary signs, and M^{3}T adds an auxiliary translation loss. SignSparK (Low et al., 2026) shows that a diffusion generator becomes strong when given ground-truth keyframes of the test clip and falls to the level of content-free baselines without them, an observation about timing that anticipates ours. We build on diffusion and flow-matching SLP in continuous latents (Feng et al., 2025). None of these gloss-free systems uses an explicit lexical intermediate, whereas we show that such an intermediate is useful once it no longer depends on annotation.

Pseudo-glosses and sign spotting. LLM-derived pseudo-glosses have been used for sign language translation (Guo et al., 2026) and, with stepwise prompting, for production (An & Kawakami, 2025). These intermediates are derived from the text and inherit its order and its blindness to timing. Sign spotting locates isolated signs in continuous video, historically with mouthing or subtitle cues (Albanie et al., 2020; Momeni et al., 2020; Varol et al., 2021), and more recently by matching dictionary exemplars into continuous signing with learned embeddings (Wong et al., 2026) or joint video–text spaces (Jiang et al., 2024). These methods spot signs to annotate or recognise them and are tuned for precision on labelled benchmarks. We use spotting to condition a generator instead, so the spotter's operating point is judged by the generated signing, where missed and false anchors trade off (§4.4), and its threshold is set against a wrong-caption control rather than on annotations.

Evaluation. SLP is typically scored by a DTW-aligned joint distance to one reference (Sakoe & Chiba, 1978; Arkushin et al., 2023; Zuo et al., 2025) and by back-translation, in which a sign translator maps the generated motion back to text that is compared with the source sentence using BLEU, ROUGE or WER (Saunders et al., 2020). BackTranslation 2.0 (Cory et al., 2026) replaces the single text score with an LLM-checked pipeline of specialised tools that scores grammar, phonology, fluency and fidelity separately. MaDiS (Zuo et al., 2026) adds two learnt metrics, SiBLEU, a BLEU score between sign-tokenised generated and reference motion, and SiCLIP, a text–sign retrieval accuracy. DTW favours still motion and its warping absorbs timing errors, and SiBLEU and SiCLIP, which compare generated motion with reference motion or with embeddings learnt from it, inherit these weaknesses. Back-translation models trained on a few thousand clips reflect the phrasing of their training corpus, and their ceiling, the score on the ground truth, often goes unreported. We quantify these weaknesses with content-free baselines and measure content with a public sign translator (Li et al., 2025) read through calibrated keypoint adapters (Jin et al., 2020), reporting every score against its ceiling (§4.3).

3 METHODOLOGY

ATLAS (Fig. 1) generates the latents of a motion VAE with a flow-matching transformer (§3.1) conditioned on the sentence and on a sequence of automatically extracted timed lexical anchors (§3.3). During training, the anchors are spotted in the signing motion with a dictionary-grounded sign-identity model; at inference, where no reference motion sequence exists, they are predicted from the sentence (§3.2).

3.1 MOTION REPRESENTATION AND GENERATION

Motion VAE. Each SMPL-X frame is described by four feature streams, the body, the two hands and the face, each combining joint rotations with joint positions (for the face, expression coefficients) in a signer-centred frame (Appendix B). A VAE (Kingma & Welling, 2014) compresses these streams $4\times$ in time. Every stream has its own convolutional-transformer encoder, so that the hands, which carry most of a sign's identity, do not compete for capacity with the larger motions of the body. The streams are then merged by a fusion transformer that attends jointly over the tokens of all four, so that each stream's latent also encodes its relation to the others, such as the location of the hands relative to the body and the synchronisation of facial and manual movements. The decoder mirrors this structure. A clip of T frames yields $K{=}\lceil T/4 \rceil$ latent steps of 96 dimensions per stream, and the generator operates on the posterior means, concatenated into one sequence $z_{1:K}$ with $z_k \in \mathbb{R}^{384}$. The VAE is frozen in every experiment.

Generator. The generator is a flow-matching transformer (Lipman et al., 2023; Liu et al., 2023; Peebles & Xie, 2023) that predicts the velocity of $z_{1:K}$ and receives its condition through three inputs. (1) A *per-step condition* $c \in \mathbb{R}^{K\times d}$, with d the transformer width, is aligned with the latent sequence (§3.3); it is concatenated to the noisy latent at the input and attended to in every block. (2) A *condition memory*, the variable-length sequence of the sentence's token embeddings, and in the hybrid model also of the anchor tokens, is read by cross-attention in every block, so each latent step can retrieve any part of the sentence regardless of word order. (3) A *global vector* $g \in \mathbb{R}^d$, the mean-pooled sentence embedding plus a learned corpus embedding, summarises the whole sentence. It is added to the embedding of the flow time $t \in [0, 1]$, the position of the sample between noise and data, and modulates every block through adaptive layer normalisation (AdaLN) (Peebles & Xie, 2023). A small head predicts from it the logarithm of the output length K, which must be fixed before sampling and depends on the sentence as a whole. Sentences are encoded by the frozen encoder of the multilingual mBART-large-cc25 (Liu et al., 2020), which covers the spoken languages of all our corpora. Sampling uses classifier-free guidance (Ho & Salimans, 2022), for which the condition is replaced by learned null embeddings during training.

3.2 TIMED LEXICAL ANCHORS

Gloss conditioning considerably improves the quality of the generated signing (§4.1), but glosses require expert annotation, and approximations derived from the caption alone do not reproduce the gain, since they miss where each sign occurs (Table 2). We therefore propose to extract a partial but timed transcript from the signing itself. A *timed lexical anchor* $(w, [t_0, t_1])$ states that the dictionary sign w is performed in the time span $[t_0, t_1]$ of the clip. Anchors are extracted without continuous annotation and need not cover every sign of the clip, which matters because a dictionary covers only part of the signed vocabulary. Whatever the anchors miss can still be recovered from the sentence, which the generator receives in full (§3.3).

Anchor spotting (training). For a clip with caption s, we first propose candidate dictionary signs. The content words of s, selected by part-of-speech tagging (Honnibal et al., 2020), and its pronouns, which ASL signs by pointing, are matched to dictionary entries by lemma, surface form or multi-word expression, through a small hand-curated list of English-to-ASL aliases such as *done* $\rightarrow$ FINISH, and, failing these, through WordNet synonyms (Miller, 1995). A sign-identity encoder f embeds a motion window x, either an isolated dictionary sign or a span of continuous signing, into a shared unit-norm space, in which a sign lies close to its dictionary version whether it is performed alone or within a sentence. It is trained with an additive-margin classification loss (Deng et al., 2019) on WLASL-2000 (Li et al., 2020), and each of the 4.7k dictionary entries is represented by one prototype, the class weight for a WLASL sign or the mean embedding of its exemplars for an

ASL-Citizen sign (Desai et al., 2023). Every candidate is scored by cosine similarity against all spans of a dense grid over the clip, a dynamic programme assigns each candidate at most one span without overlaps, and candidates scoring above a threshold τ are accepted.

Two properties of f are essential. It must be sensitive to the direction of motion, because reconstruction and classification objectives match a time-reversed window as well as x itself, so a repulsion term pushes $f(x)$ away from $f(\mathrm{reverse}(x))$. Its score must also peak on the exact extent of a sign, so f scores a span in the context of the whole clip, including the states at its boundaries, and is trained to rank known spans above shifted, cropped and over-extended copies and above the best-matching grid windows that barely overlap them. The known spans come from synthetic concatenations of isolated signs and from CSL-Daily sentences, whose human-annotated glosses we align to the motion with a CTC recogniser. After a first spotting pass, f is fine-tuned once more with the anchors it accepted on the How2Sign training clips added as known spans, a round of self-training that adapts it to the co-articulation of continuous ASL.

Without labelled continuous ASL, τ is calibrated on a control that pairs each clip with the caption of another video, so that every anchor it accepts is false, and τ is set to yield $b{=}0.8$ false anchors per clip. On How2Sign this gives 2.4 anchors per clip, about two thirds of them correct, and the 377k YouTube-SL-25-ASL clips are annotated in under an hour on a single GPU. Appendix C shows examples.

Anchor prediction (inference). Spotting requires the reference motion that the generator aims to produce, so at inference the anchors are predicted from the sentence. The predictor receives the same candidates as the spotter, one *slot* per candidate keyword of the caption. A slot is represented by a learned embedding of its dictionary sign, the sign's identity prototype and the position of its keyword in the caption. A transformer encoder processes a single sequence formed by a summary token, the sentence's token embeddings and the slots, so that every slot attends to the whole sentence and to the other candidates. For every slot the predictor outputs three quantities: the probability that the spotter would accept it, the centre of its span as a distribution over normalised time, and its duration. The summary token predicts the length of the clip, which maps normalised positions to latent steps. The predictor is trained on the spotted anchors of the training clips of YouTube-SL-25 and How2Sign, so it learns from text alone which of the sentence's signs a signer produces and roughly when. At inference, slots whose presence probability exceeds a threshold, chosen on validation clips to match the spotter's anchor count, are placed at the medians of their predicted centre and duration, with at most one anchor per caption word.

3.3 Hybrid conditioning

Anchors cover only a fraction of the signs in a clip, so the condition has to state explicitly where the unidentified signs are, rather than let the few known signs spread over the whole timeline. The anchors of a clip, spotted or predicted, are written in temporal order into a token sequence, and every stretch of the timeline not covered by an anchor becomes learned `<unk>` tokens, one per 0.5 s, the approximate duration of a sign. A three-layer transformer encoder embeds this sequence, so that every token also carries information about its neighbours, and the result is tiled uniformly over the K latent steps to form c. The condition therefore states "sign X, then about two unidentified signs, then sign Y", and the number of `<unk>` tokens between two anchors fixes their relative timing. The condition memory holds the sentence tokens followed by the anchor tokens, and g, hence the predicted length, is computed from the sentence alone. During training, the sentence and the anchors are each dropped independently, in addition to the joint dropout used for guidance.

4 Experiments

Datasets. We adopt How2Sign (Duarte et al., 2021), with 28k training, 1.6k validation and 2.1k test ASL sentences, as our main benchmark for annotation-free production. For large-scale training, we add the 377k unglossed ASL sentence clips of YouTube-SL-25 (Appendix A). We also use CSL-Daily (Zhou et al., 2021), with 18k training and 1.2k test sentences annotated with human glosses, to measure what a reference lexical condition contributes and, because it contains the same sentences performed by different signers, as a reference point in the evaluation audit. All corpora are represented as SMPL-X parameters at 24 fps.

Implementation details. The motion VAE has 55M parameters and is trained on 1,560 hours of continuous signing in all 25 languages of YouTube-SL-25, on random 256-frame windows, and never sees How2Sign or CSL-Daily. Its loss combines smooth-L1 reconstruction, weighted $2\times$ on hands and face, velocity and multi-lag displacement terms, and a KL term ($\beta{=}10^{-3}$) with per-dimension free bits (Kingma et al., 2016). We train it for 100k steps with batch size 128 and AdamW under a cosine schedule from $2{\times}10^{-4}$, and scale each latent stream to unit RMS before generation. The generator has 69M parameters, 6 layers and width 512, and is trained with batch size 256 and AdamW at 10^{-4} under a cosine schedule. For classifier-free guidance, the whole condition is replaced by learned null embeddings with probability 0.15, and the hybrid model additionally drops the sentence and the anchors independently with probability 0.3 each. Generators are always trained on spotted anchors, and the predictor replaces them only at inference. When YouTube-SL-25 is added, How2Sign makes up half of every batch and the number of steps grows with the data, from 20k steps for 27k training sentences to 40k for 100k and 80k for 404k. The 404k model is the one reported as ATLAS in Table 1. Sampling integrates the flow with 25 Euler steps at guidance scale 3.0, applied to the predicted clean latent with standard-deviation rescaling at $\phi{=}0.7$ (Lin et al., 2024). The anchor predictor has 14M parameters, 6 layers and width 384. It is trained on the anchors spotted at $b{=}0.8$ in the training clips of YouTube-SL-25 and How2Sign, never seeing test clips or their motion.

Evaluation protocol. We compute DTW-PA-JPE for the body and each hand following the evaluation convention of SOKE (Zuo et al., 2025): the DTW-aligned error of SMPL-X joints after Procrustes alignment of the scored joints, with every system generating sequences of its own predicted duration. Each system is scored against the ground truth of its own SMPL-X fitting. Because DTW-PA-JPE is a path cost accumulated over the alignment, it grows with sequence length, so shorter outputs score better and a model can improve its distance simply by under-generating. We therefore also report it per frame, divided by the length of the warping path, which gives the mean joint error per matched frame pair in millimetres, and we report the ratio of output to reference duration and the motion amplitude next to the distances. Content is measured by back-translation (BT) with a state-of-the-art public pose-based translation model, Uni-Sign (Li et al., 2025), which reads the generated motion through a calibrated SMPL-X-to-keypoint adapter (Appendix E). Its score on the ground truth of a fitting is the ceiling for every model scored in that fitting, and we also report each score as a percentage of this ceiling. On CSL-Daily we additionally report the gloss WER of a frozen CTC recogniser applied to the generated motion.

4.1 Main results

Human glosses are a much stronger condition than the sentence. On CSL-Daily, with the same generator, they more than halve the gloss WER of the generated signing and bring both distances substantially closer to the reference (Table 2). Glosses derived from the caption bring almost no gain, and become useful only when the same words are placed at the true temporal positions of their signs. A derived lexical condition therefore has to supply timing, which is what timed anchors provide on How2Sign, where no glosses exist.

Table 1 compares ATLAS with prior state-of-the-art SLP approaches on How2Sign. With anchors predicted from the sentence, ATLAS reaches 2.25 BLEU-4, 21% of the ceiling of its fitting, while the same generator conditioned on the sentence alone reaches 14% and SOKE and SignSparK reach 9% of their own ceilings, close to the content-free baselines. Measured per matched frame, ATLAS is also the closest to the reference for the body and both hands. SOKE ranks first on the accumulated cost, but its outputs are shorter than the reference, with a duration ratio of 0.92, and move with only 0.79 of its amplitude, both of which lower an accumulated distance, whereas ATLAS matches the reference in duration and amplitude at 0.99. SignSparK, whose released model generates at most 304 frames, truncates 15% of the test clips and is scored against the full reference. Fig. 2 shows sequences produced by each model. ATLAS reproduces the reference hand positions and their order the closest, and its back-translation recovers more of the sentence. SOKE's hands stay close to one position, and SignSparK often moves without following the reference. Appendix D gives more examples.

Where the gain comes from. Table 4 isolates the anchors by changing only the condition given at inference. On the generator trained on How2Sign alone, anchors spotted in the reference motion raise back-translation by 36% over the sentence-only model, and withholding them from the same

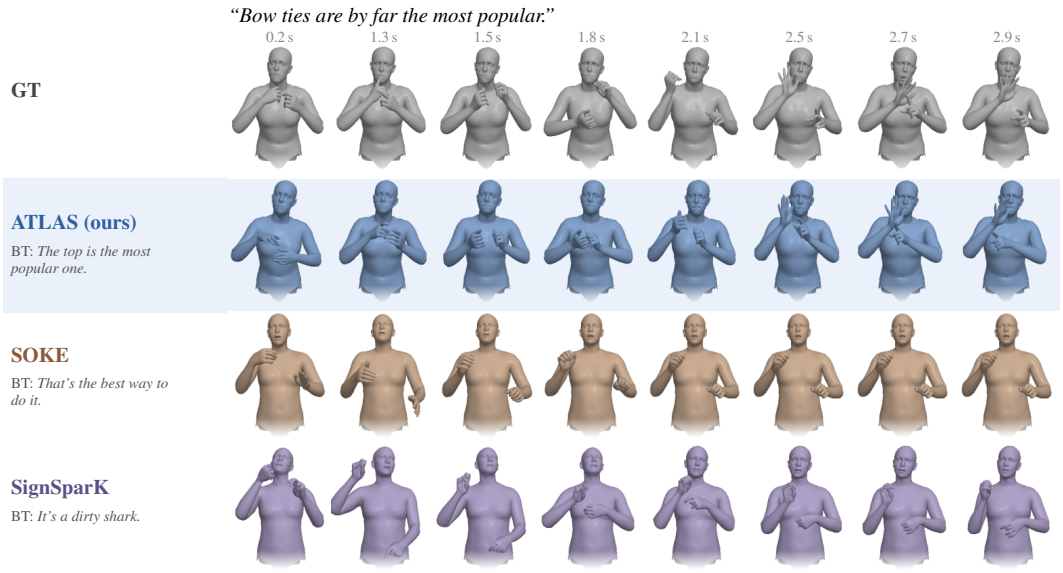


Figure 2: **Qualitative examples of sequences produced by each model.** The rows are the ground truth followed by each model's rendition, columns are aligned frames. *BT* is the Uni-Sign back-translation of each model's output.

model lowers it to the level of a still pose, so the generator relies on the anchors it was trained with. Spotting, however, requires the motion that is to be generated, so at inference ATLAS predicts the anchors from the sentence. With predicted anchors the same generator gains 50% over the sentence-only model, and the 404k generator 52%. Since predicted anchors outperform spotted ones, we checked that this gain reflects what the predictor supplies rather than information leaking from the test set. The predictor sees only the test caption and its dictionary candidates, and its anchors are not a copy of the spotter's, sharing only a third of them with a Jaccard overlap of 0.34. It selects the most sign-able content words of the sentence and places them at typical positions, whereas about a third of the spotted anchors are false detections, and mistimed anchors are harmful (§4.4). The gain also does not come from How2Sign supervision, since a predictor trained on YouTube-SL-25 alone still improves the 404k generator by 32%.

4.2 Scaling experiments

Table 5 trains the same generator, with the sentence alone and with the hybrid condition, on the 27k How2Sign training sentences and on How2Sign combined with part or all of YouTube-SL-25-ASL, for 100k and 404k sentences. Back-translation improves with data under both conditions, from 1.19 to 1.48 BLEU-4 for the sentence-only model and from 1.78 to 2.25 with predicted anchors. The anchors improve back-translation at every data size, and the anchored model trained on 27k sentences already scores above the sentence-only model trained on $14\times$ more data. Per matched frame, the distance agrees: the body error is lower at 404k than at 27k under every condition, and the anchored models are closer to the reference than the sentence-only model at every data size. The accumulated distance, in contrast, stays flat or grows slightly with data, because the duration ratio rises from 0.96 to 0.99 as the models stop under-generating, and an accumulated cost penalises longer outputs (§4.3).

4.3 Evaluation audit

Neither the distance nor back-translation is grounded in the typically used evaluation protocol, so we calibrate both with content-free baselines, transformations of the reference that carry little or no linguistic content by construction (Table 1, last block).

Table 1: **Quantitative results on How2Sign**: back-translation (BT), duration ratio (Dur.), motion amplitude (Amp.) and DTW-PA-JPE, abbreviated DTW, per matched frame and accumulated. GT rows give each fitting's BT ceiling, the third block uses the reference motion and the last block is content-free.

	BT ↑				DTW per frame (mm) ↓			DTW accumulated ↓		
System	BLEU-4	% ceil.	Dur.	Amp.	Body	Hand L	Hand R	Body	Hand L	Hand R
SOKE GT	7.3	100	1.00	1.00	0	0	0	0	0	0
SignSparK GT	7.7	100	1.00	1.00	0	0	0	0	0	0
ATLAS GT	10.6	100	1.00	1.00	0	0	0	0	0	0
NAR† (Huang et al., 2021)	–	–	–	–	–	–	–	13.94	11.80	
T2S-GPT† (Yin et al., 2024)	–	–	–	–	–	–	–	11.48	6.39	
S-MotionGPT† (Hwang et al., 2024)	–	–	–	–	–	–	–	11.23	4.39	
SOKE (Zuo et al., 2025) (no retrieval)	0.68	9	0.92	0.79	44.7	13.1	14.4	**6.93**	**2.17**	**2.38**
SignSparK (Low et al., 2026) (no keyframes)	0.67	9	0.87	1.02	44.5	14.0	13.9	9.89	3.44	3.45
ATLAS, sentence only	1.48	14	0.97	0.94	42.0	12.8	13.1	8.93	2.91	2.99
ATLAS	**2.25**	**21**	**0.99**	**0.99**	**40.5**	**12.6**	**13.0**	9.01	2.92	3.02
SOKE tokenizer round trip of GT	0.77	11	0.99	0.78	29.8	7.3	8.4	4.44	1.12	1.31
SignSparK with GT keyframes (16% of frames)	5.43	70	0.87	0.99	7.3	3.1	4.0	2.08	0.92	1.16
reference at 75% amplitude	7.84	74	1.00	0.77	5.2	2.3	2.7	1.17	0.53	0.59
baseline: still per-clip mean pose	0.75	7	1.00	0.00	22.2	10.4	12.1	4.89	2.32	2.58
baseline: reference at 25% amplitude	0.60	6	1.00	0.26	16.6	7.7	9.0	3.66	1.74	1.93
baseline: unrelated sentence	0.49	5	1.00	0.94	46.4	14.9	15.2	9.50	3.31	3.40

† Numbers as reported in the How2Sign table of SOKE (Zuo et al., 2025), not re-scored.

Table 2: **Gloss conditioning on CSL-Daily**: gloss WER and accumulated DTW-PA-JPE (DTW) for one generator.

Condition	WER ↓	DTW body ↓	DTW hands ↓
sentence	0.92	6.33	1.88
caption words	0.90	–	–
at true slots	0.75	–	–
human glosses	**0.40**	**4.74**	**1.23**

Table 3: **Spotter ablation** on How2Sign at b=0.3: number of correct anchors per clip, change over the previous row, and anchors accepted on time-reversed motion.

Spotter	# anchors ↑	Δ	Reversed ↓
synthetic sequences	1.15	–	0.27
+ CSL-Daily spans	1.22	+6%	0.74
+ span-level reversal	1.33	+8%	0.10
+ ASL-Citizen exemplars	1.40	+6%	0.09
+ self-training	**1.61**	+15%	0.03

Table 4: **Ablation of the anchor condition** on How2Sign test: BT and its change Δ over the sentence-only model of the same data size.

Condition	BLEU-4 ↑	Δ
27k sentences, condition at inference		
sentence-only model	1.19	–
predicted anchors	**1.78**	**+50%**
spotted anchors	1.62	+36%
positions scrambled	1.07	−10%
anchors withheld	0.76	−36%
27k sentences, false-anchor budget b		
b=0.3	1.35	+13%
b=0.8 (spotted anchors)	**1.62**	**+36%**
b=1.5	1.50	+26%
b=2.5	1.37	+15%
404k sentences		
sentence-only model	1.48	–
predicted anchors (ATLAS)	**2.25**	**+52%**
predictor without How2Sign	1.95	+32%
spotted anchors	1.87	+26%

A body frozen at the clip's mean pose scores better than every system on the accumulated distance, and on CSL-Daily, where sentences are also performed by a second signer, that correct rendition scores worse than the still pose, 5.32 against 4.80 body DTW-PA-JPE. Measuring the distance per frame removes the reward for short outputs but not for stillness, since the still pose has about half the per-frame error of any system.

Through Uni-Sign, every content-free baseline reads at most 7% of the ceiling, whereas the reference at 75% amplitude reads 74% and, on CSL-Daily, another signer's rendition 90%. A sign translator trained on How2Sign following Saunders et al. (2020), as is common practice, reaches only 4.25 BLEU-4 on the ground truth, of which a frozen body retains almost half. Uni-Sign is instead public and pre-trained at scale, and we read it through a keypoint adapter calibrated per fitting, which we release. The same baselines also bound published systems. SOKE's tokenizer alone reduces back-

Table 5: **Training on more data**: How2Sign test back-translation and DTW-PA-JPE, abbreviated DTW, per matched frame and accumulated over the warping path, as YouTube-SL-25-ASL is added.

			BT ↑	DTW per frame (mm) ↓			DTW accumulated ↓		
Condition	# sentences	steps	BLEU-4	Body	Hand L	Hand R	Body	Hand L	Hand R
sentence only	27k	20k	1.19	43.4	13.1	13.4	9.08	2.92	3.00
sentence only	100k	40k	1.18	42.2	13.0	13.2	8.90	2.91	2.97
sentence only	404k	80k	1.48	42.0	12.8	13.1	8.93	2.91	2.99
+ predicted anchors	27k	20k	1.78	41.0	12.6	13.3	8.70	2.84	2.98
+ predicted anchors	100k	40k	1.83	41.4	12.7	13.2	8.84	2.85	2.97
+ predicted anchors	404k	80k	**2.25**	40.5	12.6	13.0	9.01	2.92	3.02
+ spotted anchors	27k	20k	1.62	40.1	12.1	12.8	8.48	2.68	2.85
+ spotted anchors	100k	40k	1.73	40.4	12.0	12.5	8.56	2.66	2.81
+ spotted anchors	404k	80k	1.87	39.5	11.9	12.4	8.71	2.74	2.85

translation of the reference to the level of the baselines, which bounds any generator built on its tokens, and SOKE and SignSparK read at the same level unless SignSparK is given ground-truth keyframes.

Proposed protocol. Every SLP table should report three content-free baselines, the still per-clip mean pose, the reference at 25% amplitude and a length-matched unrelated sentence where available, together with the duration ratio and the distance both per frame and accumulated. Content should be measured with one shared public sign translator through a calibrated keypoint adapter and reported next to its score on the reference of the same fitting. A system whose back-translation does not exceed the baselines has produced no recoverable content, whatever its distance. Table 1 follows this protocol, and a system built on a learnt motion representation can additionally report the round trip of the reference through it, as we do for SOKE.

4.4 Ablation study

Ablation results on the anchor condition and spotter configuration are shown in Table 4 and Table 3. The anchors help only when the right sign is at the right time, since scrambling the positions of the spotted anchors lowers back-translation below the sentence-only model. Anchor precision matters as much as recall, since back-translation traces an inverted U over the false-anchor budget and peaks at b=0.8. On the spotter side, training on continuous CSL-Daily spans raises the yield, but with the reversal term applied only to whole isolated clips the spotter accepts 0.74 anchors per clip on time-reversed motion. Applying the term to spans restores direction sensitivity and raises the yield further. ASL-Citizen exemplars and one round of self-training add 6% and 15% more anchors.

5 Conclusions and Limitations

A lexical condition improves sign language production only if it conveys when each sign occurs, and such a condition can be obtained without annotation. ATLAS spots timed lexical anchors in the signing with a dictionary-grounded sign-identity model, predicts them from the sentence at inference, and combines them with the sentence in a hybrid condition. This lets it train on substantially more ASL data than prior work and reach 21% of its back-translation ceiling on How2Sign, more than twice the share of prior systems. Our audit shows that standard reference-based distance metrics reward short and still outputs, and we propose an evaluation protocol grounded in content-free baselines and a public back-translation model.

Limitations. The anchors are sparse, about 2.4 per How2Sign sentence, because a dictionary covers only part of the signed vocabulary and admitting more candidates costs precision. Spotting relies on isolated-sign dictionaries with motion exemplars, which restricts ATLAS to sign languages that have them. Finally, the SMPL-X parameters that serve as ground truth are estimated from video and contain fitting errors, which bound the quality of the generated motion.

AI USE STATEMENT

In this work, we used generative AI tools for research execution, to draft sections of the paper, and to aid and polish the writing. Specifically, AI assistants helped set up some of the experiments, drafted initial versions of some sections of the paper, and refined the writing of the manuscript. We have reviewed all AI-assisted work, including the experimental setups and every drafted or refined passage of text. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In *ECCV*, 2020.

Zhaoyi An and Rei Kawakami. Teach me sign: Stepwise prompting llm for sign language production. In *2025 IEEE International Conference on Image Processing (ICIP)*, pp. 439–444. IEEE, 2025.

Rotem Shalev Arkushin, Amit Moryossef, and Ohad Fried. Ham2Pose: Animating sign language notation into pose sequences. In *CVPR*, 2023.

Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. Neural sign actors: A diffusion model for 3d sign language production from text. In *CVPR*, 2024.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Oliver Cory, Maksym Ivashechkin, Karahan Sahin, Oline Ranum, Jianhe Low, Edward Fish, Anton Pelykh, Ozge Mercanoglu Sincan, and Richard Bowden. Backtranslation2.0 – a linguistically motivated metric to assess sign language production. *arXiv preprint arXiv:2606.28673*, 2026.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

Aashaka Desai, Lauren Berger, Fyodor Minakov, Vanessa Milan, Chinmay Singh, Kriston Pumphrey, Richard Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, and Danielle Bragg. Asl citizen: A community-sourced dataset for advancing isolated sign language recognition. In *NeurIPS Datasets and Benchmarks*, 2023.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2sign: A large-scale multimodal dataset for continuous american sign language. In *CVPR*, 2021.

Liqian Feng, Lintao Wang, Kun Hu, Dehui Kong, and Zhiyong Wang. Text2sign diffusion: A generative approach for gloss-free sign language production. In *2025 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–8. IEEE, 2025.

Jianyuan Guo, Peike Li, and Trevor Cohn. Bridging sign and spoken languages: Pseudo gloss generation for sign language translation. *Advances in Neural Information Processing Systems*, 38:77471–77499, 2026.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength natural language processing in Python. Zenodo, 2020.

Wencan Huang, Wenwen Pan, Zhou Zhao, and Qi Tian. Towards fast and high-quality sign language production. In *ACM Multimedia*, 2021.

Eui Jun Hwang, Huije Lee, and Jong C. Park. A gloss-free sign language production with discrete representation. In *FG*, 2024.

Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 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, 2024.

Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *ECCV*, 2020.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *ICLR*, 2014.

Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *NeurIPS*, 2016.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *WACV*, 2020.

Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017.

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.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *WACV*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 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.

Yunfei Liu, Lei Zhu, Lijian Lin, Ye Zhu, Ailing Zhang, and Yu Li. TEASER: Token enhanced spatial modeling for expressions reconstruction. In *International Conference on Learning Representations (ICLR)*, 2025.

Jianhe Low, Alexandre Symeonidis-Herzig, Maksym Ivashechkin, Özge Mercanoğlu Sincan, and Richard Bowden. SignSparK: Efficient multilingual sign language production via sparse keyframe learning. In *ECCV*, 2026.

George A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995.

Liliane Momeni, Gül Varol, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. Watch, read and lookup: Learning to spot signs from multiple supervisors. In *ACCV*, 2020.

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*, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

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

Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3D human pose and shape estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Progressive transformers for end-to-end sign language production. In *ECCV*, 2020.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *CVPR*, 2022.

Alexandre Symeonidis-Herzig, Jianhe Low, Ozge Mercanoglu Sincan, and Richard Bowden. M3t: Discrete multi-modal motion tokens for sign language production, 2026. URL `https://arxiv.org/abs/2603.23617`.

Garrett Tanzer and Biao Zhang. Youtube-sl-25: A large-scale, open-domain multilingual sign language parallel corpus. *arXiv preprint arXiv:2407.11144*, 2024.

Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. Read and attend: Temporal localisation in sign language videos. In *CVPR*, 2021.

Ryan Wong, Youngjoon Jang, L iliane Momeni, G ül Varol, and A ndrew Zisserman. Signmatch: Matching dictionary signs to continuous sign language video. *arXiv preprint arXiv:2609.01886*, 2026.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2023.

Aoxiong Yin, Haoyuan Li, Kai Shen, Siliang Tang, and Yueting Zhuang. T2s-gpt: Dynamic vector quantization for autoregressive sign language production from text. In *ACL*, 2024.

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

Bingfan Zhu, Biao Jiang, Sunyi Wang, Shixiang Tang, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. MotionGPT3: Human motion as a second modality. *arXiv preprint arXiv:2506.24086*, 2025.

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

Ronglai Zuo, Rolandos Alexandros Potamias, Qi Sun, Evangelos Ververas, Jiankang Deng, and Stefanos Zafeiriou. MaDiS: Taming masked diffusion language models for sign language generation. *arXiv preprint arXiv:2601.19577*, 2026.