SignGPT and the Visual Language Toolkit

Matt Brown[1], Oline Ranum[2], Edward Fish[2], Heidi Proctor[1], Bencie Woll[1], Richard Bowden[2], Kearsy Cormier[1]

[1] Deafness, Cognition, & Language Centre (DCAL), University College London,
49 Gordon Square, London WC1H 0PD
[2] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey,
Guildford, Surrey GU2 7XH
{m.brown.13, heidi.proctor.14, b.woll, k.cormier}@ucl.ac.uk
{o.ranum, edward.fish, r.bowden}@surrey.ac.uk

Abstract

SignGPT's Visual Language Toolkit (VLT) aims to remove fundamental barriers to large scale sign language modelling by developing data-driven, linguistically grounded methods for continuous sign language recognition. We first identify fundamental issues around the ecological validity of potential data sources (e.g. broadcast media with interpreted signing or captions, scraping of social media). We contrast these with the currently highly resource-intensive development of curated sign language corpora based on linguistic principles. The VLT addresses this scarcity of high quality sign language data by providing semi-automated glossing and other recognition tools, driving large scale corpus expansion without sacrificing linguistic principles. Unlike prior systems that rely on sparse glossing, the project integrates dense temporal annotation, non-manual and non-lexical feature tracking, and transformer-based architectures to capture the multimodal and spatial structure of signing. By aligning machine vision innovation with linguistic insights and community-embedded evaluation, SignGPT establishes a foundation for robust and extensible sign language models.

Keywords: sign language linguistics, large language models, glossing, transformers, corpus linguistics

1. Overview

SignGPT (https://sites.google.com/view/signgpt) is a 5-year grant from the Engineering and Physical Sciences Research Council, part of UK Research and Innovation. The project brings together specialists in machine vision, generative AI, and sign language linguistics from Surrey University, Oxford University, and University College London. It aims to develop the first large-scale, end-to-end AI system capable of bidirectional translation between British Sign Language (BSL) and spoken/written English. To that end, we treat languages in the visual-spatial modality as deserving of their own large language models (LLMs) rather than as derivatives of a spoken language. Unlike prior systems that have relied on limited vocabularies and rule-based translation, the project integrates computer vision, sign linguistics, and generative AI to address the full complexity of continuous signing, including non-manual features and spatial grammar. The long term ambition is to create a conversational sign language system analogous to ChatGPT, allowing deaf users to interact naturally in sign and receive fluent, photorealistic signed responses.

Training LLMs and other foundational models requires a large amount of data. Sign languages do not have a written form, so historically the creation of suitable datasets has been hampered by the intensive labour required for accurate and exhaustive linguistic annotation. For instance, only about 15% of the video data in the BSL Corpus (Schembri et al., 2014) has been annotated at all to date. To address this issue, SignGPT is developing semi-automatic annotation tools that reduce dependence on costly manual glossing while still preserving linguistic validity. This Visual Language Toolkit (VLT) will be released openly, benefiting the wider research community. The project itself will use these tools to curate and annotate a massive sign language dataset, expanding existing annotated BSL corpora by an order of magnitude.

On the recognition side, the project moves beyond isolated sign recognition to dense temporal gloss annotation, i.e. comprehensive time-aligned labelling which includes non-manual and non-lexical material, rather than only representing manual forms from the core lexicon. We combine skeletal uplift, facial landmark tracking, and transformer-based models that are robust enough to handle co-articulation and signer variation. On the production side, the project avoids the traditional use of animated avatars, favouring photorealistic digital humans generated using diffusion models, volumetric capture, and differentiable rendering. Another key innovation is the use of self-supervised learning to build predictive language models directly from sign representations. Rather than relying solely on glossing as a proxy for language, the project is exploring how symbolic representations at the lexical, sub-lexical, and gestural levels can support large-scale modelling of sign structure and meaning.

2. The data problem

As noted above, a fundamental linguistic challenge addressed by the project is that sign languages do not have a standardised written form: they cannot be reduced to sequences of discrete "words". As a result, standard Natural Language Processing (NLP) assumptions based on text analysis around linear order, tokenisation, and stable form-meaning mappings do not necessarily hold. Natural continuous signing integrates both lexical and non-lexical elements: just ~60% of spontaneous BSL consists of lexical signs (Fenlon et al. 2014). The remainder consists largely of productive constructions that exploit 3D space, iconicity, and context. These include manual forms such as pointing, depicting signs, and enactment, as well as non-manual features such as brow and eye movements, mouthings, mouth gestures, head movements, and body postures. These elements interact simultaneously rather than sequentially, with meaning distributed across multiple articulators, and are thus not easily symbolised or glossed. The project addresses this by explicitly modelling non-lexical structures, rather than treating them as noise or ignoring them. Linguistic analysis of indexing, depiction, and enactment (following e.g. Ferrara and Hodge, 2018) feeds directly into the translation models, ensuring that these crucial features of spatial grammar and reference tracking unique to sign languages are incorporated.

2.1 Current state of the Sign Language Processing (SLP) field

Interest in the computational modelling of sign languages has accelerated over the last decade, with SLP research increasingly represented at top-tier machine learning conferences (e.g. Camgoz et al., 2018; Jang et al., 2024; Li et al., 2025; Zuo et al., 2025; Fish & Bowden, 2025). Despite the breadth of modeling challenges, research has largely concentrated on the three primary tasks - recognition, production, and translation - typically operating on video or coarse pose data together with gloss annotations or free-text translations (Bragg et al., 2019; Desai et al., 2024; Fox et al., 2025).

Sign-to-text translation is commonly framed as Sign Language Translation (SLT), encompassing both Sign Language Recognition (SLR, meaning sign-to-gloss) and translation to spoken language (either gloss-to-text or sign-to-text). Historically, these components have often been treated as separate modelling problems (Zhou et al., 2021a; Zhou et al., 2021b; Yin et al., 2021; Chen et al., 2022; Chen et al., 2022; Zhang et al., 2023), organised as a pipeline from recognition to translation and frequently without explicit attention to distinctions between lexical and non-lexical structures or broader linguistic organisation (De Coster et al., 2024). More recent work has introduced gloss-free approaches that model sign-to-text mapping end-to-end rather than via gloss intermediates (Zhou et al., 2023; Sincan et al., 2025; Wong et al., 2024; Chen et al., 2024; Asasi et al., 2025; Chen et al., 2025; Kim et al., 2025). However, both gloss-based and gloss-free paradigms still tend to prioritise mapping coarse manual features to spoken language text, often without modelling the full compositional and multimodal structure of signed communication (De Coster et al., 2024). Text-to-sign research similarly tends to prioritise surface-level appearance and reconstruction, with limited evaluation of linguistic composition (e.g. inclusion of productive constructions) or systematic incorporation of non-manual features (as noted by Bragg et al., 2019). This misunderstanding is reflected in the fact that text-to-sign research is often framed as Sign Language Production or Sign Language Generation, when actually it is translation to the same as extent as sign-to-text, just in the opposite direction.

Given the low-resource nature of sign language data compared to spoken languages, advancing the field requires methodologies and evaluation metrics that explicitly assess the extent to which models capture more complex visual linguistic functions and non-lexical language components. The following sections outline key challenges and research directions for broadening SLP beyond current task formulations towards modelling full, naturalistic, spontaneous, and conversational signing.

2.2 Influence and the use of available sign language data repositories

The development of the SLP field has been strongly shaped by the availability and characteristics of existing datasets, which for continuous signing largely consist of interpreted broadcasts (e.g. Camgoz et al., 2018; Albanie et al., 2021; Li et al., 2025), content sourced from social media (e.g. Shi et al., 2022; Uthus et al., 2023; Tanzer & Zhang, 2024) or curated linguistic corpora (e.g. Crasborn & Zwitserlood, 2008; Schembri et al., 2014; Konrad et al., 2020). Because curated corpora are difficult to scale and linguistic annotation is costly, interpreted and online sources are frequently used for training and benchmarking models (De Coster et al., 2024). For most of these benchmarks, annotations are typically limited to weakly aligned captions or gloss-level labels.

Reliance on such data can bias modelling priorities. Interpreted signing produced under live broadcast constraints differs from everyday interaction. Simultaneous interpreting and sight translation are cognitively demanding even for experienced practitioners regardless of modality, and the broadcast medium itself may limit the use of repair strategies typical in natural conversation. Interpreted material is therefore quite likely to contain errors (e.g. omissions, substitutions), flattening of propositional structure, explicitations (overt target text statements or connectives that were merely implied in the source), and the over-use of formulaic sequences (Ding 2017; Tang &

Li, 2017; Gumul, 2021; Li & Halverson, 2022; Huang et al., 2023; Wang, 2025). There may also be a detectable difference between hearing and deaf interpreters with regard to sign language target text design: Stone (2009) suggests that the former are more clearly influenced by source text structure, while the latter are more likely to shape their texts as audience-focused renditions. For non-interpreted spontaneous sign language data, the use of material sourced from the internet raises ethical concerns around consent and copyright, and it usually lacks reliable information about signer proficiency or context (Fox et al., 2024; De Coster et al., 2024). Finally, broadcast captions are not necessarily verbatim representations of speech: principled decisions are routinely made around omissions and simplifications due to limitations of the medium such as available screen estate and reading speeds (Schilperoord et al., 2005; Romero-Fresco, 2016, 2020). As a result, current datasets provide only limited coverage of natural, spontaneous signing, potentially shaping models toward features that diverge from those needed for real-world applications. Within this broader context, the SignGPT project contributes through the VLT to support scalable annotation of naturalistic signing data and linguistically grounded corpora.

2.3 Challenges in the use of gloss annotations

From a modelling perspective, the reliance on glosses as an intermediate representation is both useful and limiting. Polysemy and variation present particular linguistic challenges. A single sign form can represent multiple meanings, depending on context. The BSL lexicon also exhibits sociolinguistic variation (Stamp et al. 2014). The project mitigates this by distinguishing contextual glosses from ID glosses, supporting lemmatisation that groups together semantically and phonologically related variants while preserving contextual interpretation (Fenlon et al., 2015). Transformer-based models are then trained to resolve meaning using discourse-level context and non-manual information rather than simply relying on one-to-one sign/word mappings. Translation quality will be evaluated not only via automatic metrics but also through expert human judgement by qualified deaf translators, reflecting linguistic adequacy and naturalness rather than surface similarity.

2.4 Constituent order, simultaneity, and linear modelling paradigms

The reliance on gloss annotations also highlights a broader issue: the misalignment between simultaneity and organisation in sign languages and the largely linear, sequential modelling paradigms inherited from spoken-language NLP (De Coster et al., 2024). While all natural languages exhibit ordering preferences, grammatical relations in sign languages are generally less dependent on strict linear sequencing and are comparatively flexible (Baker, 2016). Consequently, signed grammatical structures may map onto multiple spoken-language sequences, and vice versa, a variability not well captured by traditional n-gram evaluation metrics such as BLEU (Papineni et al., 2002).

Linearised intermediate representations, including gloss strings or caption-aligned tokens, provide practical supervision but introduce an information bottleneck that can obscure spatial grammar, productive constructions, and flexible constituent ordering. Progress in SLP will likely require modelling approaches that move beyond strictly sequential assumptions toward representations capable of capturing multimodality, abstract use of space, cross-articulator temporal alignment, and inductive biases informing productive language use.

2.5 Inclusion of non-lexical material as input features

A related challenge concerns the representation of non-lexical material itself. Non-manual signals can contribute to lexical, grammatical, prosodic, and discourse-level meanings (Pfau & Quer, 2010). Despite their fundamental role, many pose-based SLP approaches focus primarily on coarse pose and motion (De Coster et al., 2024), often neglecting these cues. Although still relatively limited, work incorporating non-manual features into model inputs has demonstrated improved performance (Zheng et al., 2021; Dey et al., 2022; Miranda et al., 2022). Productive constructions such as depicting signs, spatial indexing, classifier constructions, and prosodic modulation likewise encode information that is only weakly, if at all, captured in gloss-based annotations, where they are frequently reduced to coarse categories.

From a modelling perspective, this necessitates multimodal representations that extend beyond handshape, pose trajectories, or low-information RGB features to include non-manual articulators, spatial relationships, and dependency patterns. Many grammatical functions, including directionality, clause type marking and adverbial modification, are conveyed through spatial and articulatory modulation rather than discrete lexical items, while constructions such as depicting signs encode meaning through the interaction of handshape, movement, and space (Baker, 2016). Without these dimensions, models risk privileging lexical recognition over broader linguistic competence.

2.6 Towards real-world generalisation

To improve real-world applicability, SLP must move towards greater use of fluent deaf conversational data, deaf-presented content, and ethically sourced naturalistic signing, supported by scalable annotation tools such as the VLT. More broadly, progress depends not only on dataset scale but on modelling and evaluation approaches that reflect the multimodal, spatial, and partially non-lexical structure of sign language. Within this context, the SignGPT project develops a conversational agent incorporating grounded translation pipelines, both lexicalised and productive language structures, and attention to linguistic diversity. Addressing these factors is essential for systems that generalise beyond constrained benchmarks toward real-world interaction, motivating the requirements outlined in the following section.

3. Visual Language Toolkit (VLT)

The accessibility of tools for machine translation and processing of sign language remains limited, with users facing multiple technical barriers when attempting to deploy existing methods for downstream applications. These barriers include poorly documented code, large and fragile dependency chains, dataset-specific implementations, and computational requirements that place many approaches out of reach for researchers without access to high-performance infrastructure. Compounding this, the majority of research in this area is driven primarily by publication incentives, and consequently few studies have released their methodologies in forms that are readily usable by the broader communities who stand to benefit, including linguists, deaf communities, educators, and researchers in adjacent fields.

The Visual Language Toolkit (VLT) is designed to address these limitations, providing a modular suite of software tools for the annotation, analysis, modelling, and generation of visual languages, with an initial focus on sign languages. Developed within the SignGPT programme and associated research funded by Google.org, the VLT will provide the technical infrastructure needed to bridge continuous sign language video, linguistic representations, and large-scale machine learning models. This effort is inspired in part by toolkits that have proven highly successful in adjacent fields, such as NLTK for natural language processing (Bird, 2006) and SpeechBrain for speech recognition (Ravanelli et al., 2024), which demonstrate the value of unifying complex ML pipelines into accessible, well-documented frameworks.

At its core, the VLT will integrate tools for automatic and semi-automatic annotation of sign language video, encompassing dense temporal gloss annotation, sign spotting, phrase and sentence segmentation, signer and video anonymisation, skeletal estimation and correction, and sign–text alignment. For machine learning researchers, the toolkit will additionally provide a dedicated Python library and API for accessing datasets, extracting features from state-of-the-art representation models, and computing a unified set of validated metrics for evaluating translation and production results. A central design principle is interoperability with established linguistic annotation platforms such as ELAN, enabling linguists and sign language researchers to work with this data at scale without necessarily adopting additional software. Early development has focused on integrating state-of-the-art methods for sign segmentation, sign spotting, automatic pseudo-gloss generation, and ID-glossing, which we describe in the following subsections.

3.1 Segmentation Tool

Segmentation is concerned with the identification of temporal boundaries within continuous signing (Ormel and Crasborn 2012). This is a challenging problem both for humans and automated systems since no single cue can reliably predict boundary locations. For example, boundaries can be signalled by multiple combinations of manual and non-manual markers. Cues relevant to segmentation at the lexical level might include phonological and prosodic transitions (Brentari, 1998); at the intonational phrase level they might include eye blinks, pauses, changes in head position and eye gaze direction, and/or shifts in body posture (Ormel and Crasborn, 2012). Perception studies have shown that while signers are able to segment discourse more consistently than non-signers, non-signers do demonstrate some sensitivity to visually salient cues such as pauses and hand drops, suggesting that at least some boundary markers are accessible independently of linguistic knowledge (Fenlon et al., 2007; Brentari et al., 2011). Early work on automatic boundary detection in signed languages has explored video-based feature extraction of facial and manual cues (Piater et al., 2010; Jantunen et al., 2010), though robust automatic segmentation remains an open challenge given the multi-articulatory nature of sign and the absence of any single dominant predictor (Ormel and Crasborn, 2012). Despite these difficulties, more recent work has demonstrated that machine learning systems can predict human-annotated boundaries on certain datasets with up to 85% accuracy using skeletal pose and hand data alone (He et al., 2025).

Within the VLT, we incorporate the state-of-the-art segmentation approach presented by He et al. (2025), which annotates frames with Begin, In, and Out (BIO) labels and employs a self-attention mechanism to learn, at each timestep, how to optimally combine body and hand features for boundary prediction. Figure 1 illustrates representative outputs on a sample from the MeinDGS dataset (Konrad et al., 2020), showing predictions and ground-truth, including successful segmentations alongside cases of over- and under-segmentation: the method performs well considering the ambiguous nature of signing boundaries. The VLT simplifies this pipeline so that users can upload video directly and receive

ELAN-compatible annotation files with proposed boundaries for review. A browser-based visualisation tool is also provided for interactive inspection of segmentation outputs.

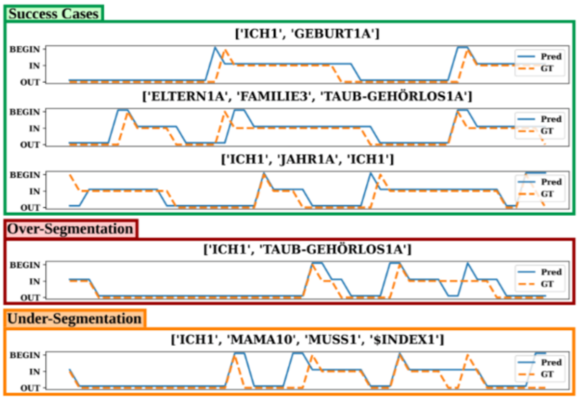


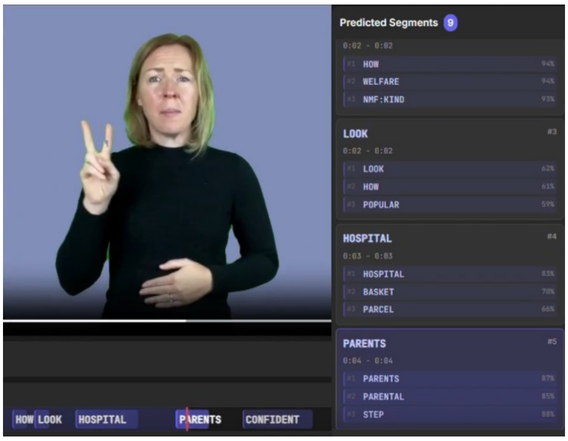
Figure 1: Examples of segmentation accuracy on a sample from the MeinDGS dataset.

3.2 Sign spotting tool

Sign spotting is the task of identifying specific lexical signs or other annotations within continuous signing by comparing sequences against a library of isolated examples drawn from lexical databases or dictionaries. By grounding spotted instances in verified reference examples, this process can be used to expand the coverage and diversity of sign dictionaries or to increase the annotation density of existing corpora.

The VLT provides an interface and API for sign spotting built upon SignRep (Wong, Camgöz, and Bowden, 2024), a state-of-the-art sign language representation model. SignRep is a masked autoencoder trained to reconstruct skeletal pose estimations from video. Through this reconstruction objective, the model learns general-purpose representations of sign segments within a compressed latent space that can be compared using cosine similarity or other distance metrics. Crucially, because the network operates on reconstructed skeletal features rather than raw RGB input, the resulting embeddings are substantially less sensitive to signer appearance, mitigating a well-known source of bias in sign recognition systems.

The spotting tool is exposed via an API that enables comparison of these learned embeddings against multiple reference dictionaries. Results can be exported in ELAN format, facilitating evaluation of inter-annotator agreement or analysis of production variation across signers. Importantly, the toolkit is designed so that the segmentation pipeline described in Section 3.1 can be composed directly with the spotting module: users can first extract temporal segments from continuous video, then perform spotting over those segments using pre-extracted dictionary embeddings provided as part of the toolkit. Figure 2 illustrates this combined pipeline on an example sequence. First, continuous signs are segmented using the method described in section 3.1. Then we embed these segments as representations using the method described above and compare with the most likely sign candidate in a dictionary of isolated signs. In the top right we can see the English source text that was translated into BSL by the signer. In the bar at the bottom are the sign segments and predicted signs. On the right are other possible candidate signs in the dictionary.


Predicted Segments 9

0:02 - 0:02
#1 HOW 94%
#2 WELFARE 94%
#3 NMF:KIND 93%

LOOK #3
0:02 - 0:02
#1 LOOK 62%
#2 HOW 61%
#3 POPULAR 59%

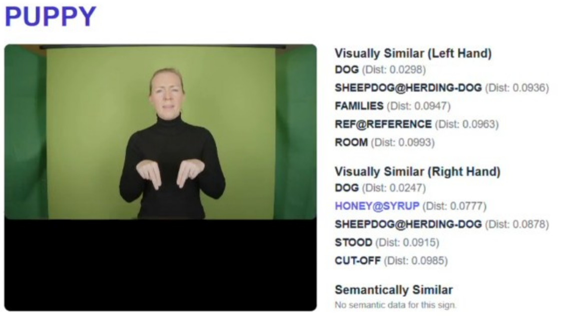
HOSPITAL #4
0:03 - 0:03
#1 HOSPITAL 83%
#2 BASKET 78%
#3 PARCEL 66%

PARENTS #5
0:04 - 0:04
#1 PARENTS 87%
#2 PARENTAL 85%
#3 STEP 88%

Figure 2: The sign segmenter and sign spotter interface.

3.3 Other tools

Additional tools within the VLT leverage pre-trained models for 3D body pose estimation (Loper et al., 2023) and hand reconstruction (Pavlakos et al., 2024) to generate high-quality mesh representations of signers from video. These reconstructions serve multiple purposes within the toolkit, from feature extraction and linguistic analysis to appearance transfer and signer anonymisation.


PUPPY

Visually Similar (Left Hand)
DOG (Dist: 0.0298)
SHEEPDOG@HERDING-DOG (Dist: 0.0936)
FAMILIES (Dist: 0.0947)
REF@REFERENCE (Dist: 0.0963)
ROOM (Dist: 0.0993)

Visually Similar (Right Hand)
DOG (Dist: 0.0247)
HONEY@SYRUP (Dist: 0.0777)
SHEEPDOG@HERDING-DOG (Dist: 0.0878)
STOOD (Dist: 0.0915)
CUT-OFF (Dist: 0.0985)

Semantically Similar
No semantic data for this sign.

Figure 3: Example of a heuristic method for comparing hand shapes extracted using the HaMER framework (Pavlakos et al., 2024)

The extracted 3D hand pose keypoints can be used independently for linguistic analysis. In Figure 3, we demonstrate an experiment in which isolated hand keypoints, without spatial or temporal context, are compared against a library of isolated sign examples to identify hand shape similarities across signs. The results reveal expected correspondences, such as between puppy and dog, but also surface less obvious relationships: for instance, honey appears among the nearest neighbours because, in BSL, the dominant hand of this sign employs the same handshape. This capability has direct applications in lexicographic analysis and dictionary construction.

Figure 4 illustrates the mesh reconstruction pipeline: keypoint and pose estimations are extracted for various articulators (body skeleton, hand keypoints, head pose) using multiple methods according to their relative limitations and strengths. Building on these reconstructions, Figure 5 demonstrates how SMPL body meshes (Loper et al., 2023) can be extracted from signing video and remapped to novel identities using Gaussian splatting (Zhang et al., 2025). First, the signer's appearance is separated from the underlying pose, normalised, and then this can be remapped to another identity. Signs from isolated dictionaries can then be stitched together to form smooth sequences of signs using techniques such as flow-matching (Lipman et al., 2022). This remapping has direct applications in signer anonymisation as well as in driving human avatars for sign language production.

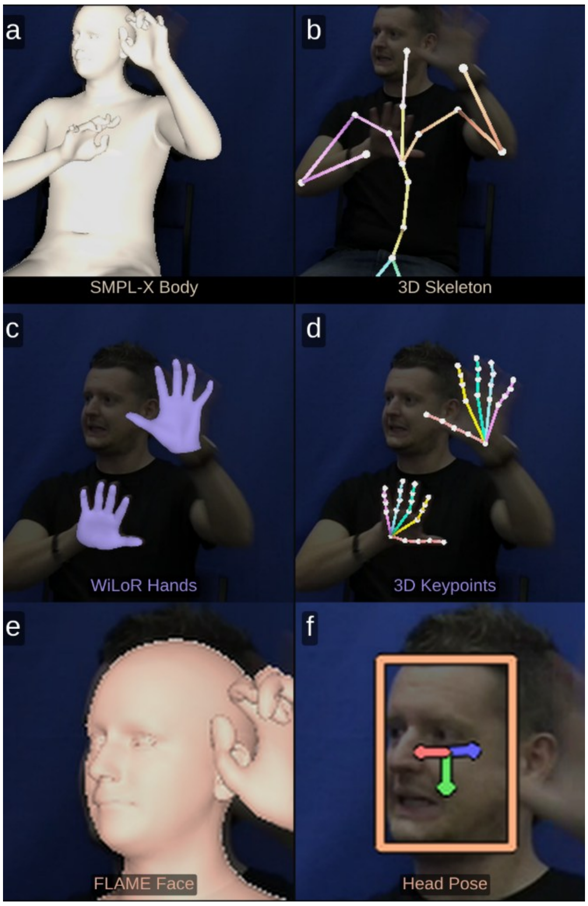


Figure 4: Example of extraction pipeline via SMPL-X (Pavlakos et al., 2019) with 3D keypoints for hands via WiLoR (Potamias et al, 2025) and head pose via FLAME (Li et al., 2017)

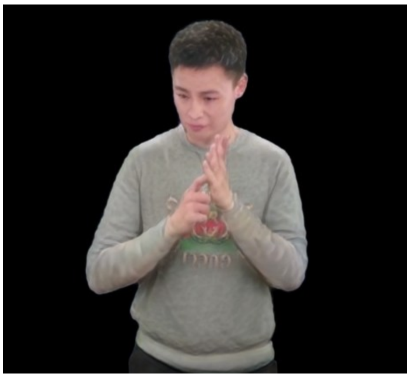


Figure 5: Example of “re-skinning” multiple different signers into a continuous sequence with a new appearance using SMPL and Guava

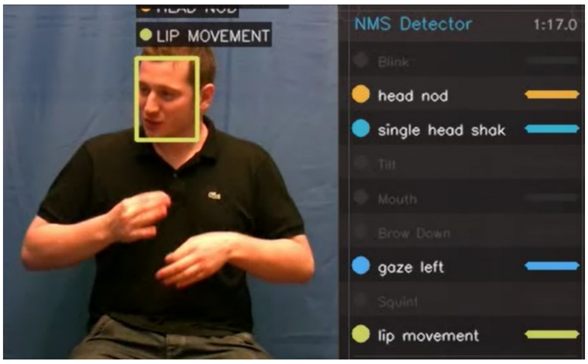


Figure 6: Example of predicted non-manual articulations based on keypoints extracted from the method described in Liu et al. (2025)

Additional tools include heuristic based algorithms that use 3D face estimation features extracted using the method “Teaser” (Liu et al., 2025) to predict when a signer is blinking, tilting their head, nodding, shaking their head, along with methods for gaze detection and eye-brow movement. As shown in Figure 6, these can be visualised over the videos or returned as ELAN files to the user.

As research in 3D reconstruction and visual representation continues to advance, the VLT will incorporate and standardise emerging methods within its modular architecture, following the development framework outlined by Ravanelli et al. (2024).

3.4 Uses for the VLT

Beyond annotation, the VLT aims to support the full sign language processing pipeline. It will incorporate technologies for continuous translation between signed and spoken languages, providing interfaces between visual representations, symbolic linguistic units, and large language models. Although developed primarily for sign languages, the toolkit is explicitly designed to generalise to other forms of visual and multimodal communication, including co-speech gesture in interaction.

The VLT is intended both as a research platform and as a community-facing resource.

Components of the toolkit are to be released as open-source software to support reuse by the wider research community, while selected capabilities will be exposed through web-based demonstrations and applications. These may include dictionary search tools, sign language learning and assessment tools, signer personalisation, and conversational interfaces that connect sign language input and output to large language models. We hope that through this dual role, the VLT will underpin advances in sign linguistics, computer vision, and machine learning, while also enabling practical, user-driven applications co-created with deaf communities (Fox et al., 2023; Desai et al., 2024).

4. Summary

We have presented SignGPT's Visual Language Toolkit project as a transformative step towards genuinely accessible sign language AI, directly addressing the linguistic and technological barriers that have historically limited progress in this field. Rather than treating sign languages as a simple sequence of glosses or as representations of spoken language, the work recognises their multimodal, spatial, and simultaneous nature with its approach to handling non-lexical signs, non-manual features, and other forms of context. By developing robust semi-automatic glossing tools that support large scale data expansion, as well as taking a community-embedded design and evaluation approach to transformer-based modelling, we aim to overcome the bottlenecks of data scarcity and linguistic over-simplification. Through this integration of linguistic insight and machine learning innovation, SignGPT establishes a foundation for robust and extensible sign language models capable of supporting real-world communication needs while respecting the linguistic integrity of BSL and other sign languages.

5. Acknowledgements

This work is supported by the UK Engineering and Physical Sciences Research Council [EP/Z535370/1], and by funding from Google.org via the AI for Global Goals scheme. We also acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR). Isambard-AI is operated by the University of Bristol and is funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023].

6. Bibliographical References

Albanie, S., Varol, G., Momeni, L., Bull, H., Afouras, T., Chowdhury, H., Fox, N., Woll, B., Cooper, R., McParland, A., et al. (2021). BBC-Oxford British Sign Language dataset. arXiv preprint arXiv:2111.03635.

Asasi, S., Lakhal, M.I., Sincan, O.M., & Bowden, R. (2025). Beyond gloss: A hand-centric framework for gloss-free sign language translation. In Proceedings of the 36th British Machine Vision Conference (BMVC 2025).

Baker, A. (2016). The linguistics of sign languages: An introduction (1st ed.). John Benjamins Publishing Company.

Bird, S. (2006). NLTK: the natural language toolkit. Proceedings of the COLING/ACL 2006 interactive presentation sessions.

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., Vogler, C., & Morris, M. R. (2019). Sign language recognition, generation, and translation: An interdisciplinary perspective. In Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)

Brentari, D. (1998). A prosodic model of sign language phonology. Cambridge, MA: MIT Press.

Brentari, D., González, C., Seidl, A., and Wilbur, R. (2011). Sensitivity to visual prosodic cues in signers and nonsigners. Language and Speech, 54(1), 49–72.

Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., & Bowden, R. (2018). Neural sign language translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Chen, Y., Wei, F., Sun, X., Wu, Z., & Lin, S. (2022). A simple multi-modality transfer learning baseline for sign language translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Chen, Y., Zuo, R., Wei, F., Wu, Y., Liu, S., & Mak, B. (2022). Two-stream network for sign language recognition and translation. Advances in Neural Information Processing Systems.

Chen, Z., Zhou, B., Huang, Y., Wan, J., Hu, Y., Shi, H., Liang, Y., Lei, Z., & Zhang, D. (2025). C^{2}RL: Content and context representation learning for gloss-free sign language translation and retrieval. IEEE Transactions on Circuits and Systems for Video Technology.

Chen, Z., Zhou, B., Li, J., Wan, J., Lei, Z., Jiang, N., Lu, Q., & Zhao, G. (2024). Factorized learning assisted with large language model for gloss-free sign language translation. In Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 7071–7081).

De Coster, M., Shterionov, D., Van Herreweghe, M. et al. (2024). Machine translation from signed to spoken languages: state of the art and challenges. Univ Access Inf Soc 23, 1305–1331

Desai, A., De Meulder, M., Hochgesang, J. A., Kocab, A., & Lu, A. X. (2024). Systemic Biases in Sign Language AI Research: A Deaf-Led Call to Reevaluate Research Agendas. In E. Efthimiou, S.-E. Fotinea, T. Hanke, J. A. Hochgesang, J. Mesch, & M. Schulder, Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources Torino, Italia.

Dey, S., Pal, A., Chaabani, C., & Koller, O. (2022). Clean text and full-body transformer: Microsoft's submission to the WMT22 shared task on sign language translation. In Proceedings of the Seventh Conference on Machine Translation. Association for Computational Linguistics.

Ding, Y. (2017). Using Propositional Analysis to Assess Interpreting Quality. International Journal of Interpreter Education, 9(1).

Fenlon, J., Cormier, K., & Schembri, A. (2015). Building BSL SignBank: The lemma dilemma revisited. International Journal of Lexicography, 28(2), 169-206.

Fenlon, J., Denmark, T., Campbell, R., and Woll, B. (2007). Seeing sentence boundaries. Sign Language and Linguistics, 10(2), 177–200.

Ferrara, L., & Hodge, G. (2018). Language as description, indication, and depiction. Frontiers in Psychology, 9:716.

Fish, E., & Bowden, R. (2025). Geo-Sign: Hyperbolic contrastive regularisation for geometrically aware sign language translation. In Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems.

Fox, N., Woll, B., & Cormier, K. (2023). Best practices for sign language technology research. Universal Access in the Information Society, 24, 69-77.

Gumul, E. (2021). Explicitation and cognitive load in simultaneous interpreting. Interpreting. International Journal of Research and Practice in Interpreting, 23(1), 45–75.

He, Low Jian, et al. (2025). Hands-on: Segmenting individual signs from continuous sequences. IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG), IEEE, 2025.

Huang, D. F., Li, F., & Guo, H. (2023). Chunking in simultaneous interpreting: the impact of task complexity and translation directionality on lexical bundles. Frontiers in Psychology, 14, 1252238.

Jang, Y., Raajesh, H., Momeni, L., Varol, G., & Zisserman, A. (2025). Lost in translation, found in context: Sign language translation with contextual cues. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Jantunen, T.J., Koskela, M., Laaksonen, J.T., and Raino, P.I. (2010). Towards the automatic visualization and analysis of signed language prosody: Method and linguistic issues. In Proceedings of the 5th International Conference on Speech Prosody, May 11–14, Chicago.

Kim, J., Jeon, H., Bae, J., & Kim, H. Y. (2025). Leveraging the power of MLLMs for gloss-free sign language translation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 21048–21058).*

Konrad, R., Hanke, T., Langer, G., Blanck, D., Bleicken, J., Hofmann, I., Jeziorski, O., König, L., König, S., Nishio, R., Regen, A., Salden, U., Wagner, S., Worseck, S., Böse, O., Jahn, E., Schulder, M. (2020). MEINE DGS – annotiert. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release / MY DGS – annotated. Public Corpus of German Sign Language, 3rd release [Dataset]. Universität Hamburg.

Li, T., Bolkart, T., Black, M. J., Li, H., & Romero, J. (2017). Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, *36*(6), 17. Association for Computing Machinery.

Li, Y., & Halverson, S. L. (2022). Lexical bundles in formulaic interpreting: A corpus-based descriptive exploration. Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association, 19(1), 33–56.

Li, Z., Zhou, W., Zhao, W., Wu, K., Hu, H., & Li, H. (2025). Uni-Sign: Toward unified sign language understanding at scale. In Proceedings of the Thirteenth International Conference on Learning Representations.

Loper, Matthew, et al. (2023). SMPL: A skinned multi-person linear model. Seminal Graphics Papers: Pushing the Boundaries, Volume 2, 851-866.

McIntosh-Smith, S., Alam S. R. and Woods, C. (2024). Isambard-AI: a leadership class supercomputer optimised specifically for Artificial Intelligence.

Miranda, P. B., Casadei, V., Silva, E., Silva, J., Alves, M., Severo, M., & Freitas, J. P. (2022). TSPNet-HF: A hand/face TSPNet method for sign language translation. In Proceedings of the Ibero-American Conference on Artificial Intelligence. Springer.

Ormel, E., and Crasborn, O. (2012). Prosodic Correlates of Sentences in Signed Languages: A Literature Review and Suggestions for New Types of Studies. Sign Language Studies, vol. 12, no. 2, pp. 279–315, JSTOR.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, & D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics.

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., & Black, M. J. (2019). Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern*

Recognition, *2019-June*, 10967–10977.

Pavlakos, G. et al. (2024). Reconstructing hands in 3d with transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Pfau, R., & Quer, J. (2010). Nonmanuals: their grammatical and prosodic roles. In D. Brentari (Ed.), Sign Languages (pp. 381–402). Cambridge: Cambridge University Press.

Piater, J., Hoyoux, T., and Du, W. (2010). Video analysis for continuous sign language recognition. In P. Dreuw et al. (Eds.), Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, 192–195. Paris: ELRA.

Potamias, R.A., et al. (2025). Wilor: End-to-end 3D hand localization and reconstruction in-the-wild. Proceedings of the Computer Vision and Pattern Recognition Conference.

Ravanelli, M., Parcollet, T., Moumen, A., de Langen, S., Subakan, C., Plantinga, P., … Esteve, Y. (2024). Open-Source Conversational AI with SpeechBrain 1.0. Journal of Machine Learning Research, 25.

Romero, J., Tzionas, D., & Black, M. J. (2017). Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, *36*(6), 245:1--245:17.

Romero-Fresco, P. (2016). Accessing communication: The quality of live subtitles in the UK. Language & Communication, 49(4), 56–69.

Romero-Fresco, P. (2020). Negotiating quality assessment in media accessibility: the case of live subtitling. Universal Access in the Information Society 2020 20:4, 20(4), 741–751.

Schilperoord, J., de Groot, V., & van Son, N. (2005). Nonverbatim captioning in Dutch television programs: A text linguistic approach. Journal of Deaf Studies and Deaf Education, 10(4), 402–416.

Shi, B., Brentari, D., Shakhnarovich, G., & Livescu, K. (2022). Open-domain sign language translation learned from online video. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

Sincan, O. M., Low, J. H., Asasi, S., & Bowden, R. (2025). Gloss-free sign language translation: An unbiased evaluation of progress in the field. Computer Vision and Image Understanding, 261, 104498.

Stamp, R., Schembri, A., Fenlon, J., Rentelis, R., Woll, B., & Cormier, K. (2014). Lexical variation and change in British Sign Language. PLoS ONE, 9(4), e94053.

Stone, C. (2009). Toward a Deaf Translation Norm. Gallaudet University Press.

Tang, F., & Li, D. (2017). A corpus-based investigation of explicitation patterns between professional and student interpreters in Chinese-English consecutive interpreting. Interpreter and Translator Trainer, 11(4), 373–395.

Tanzer, G., & Zhang, B. (2024). YouTube-SL-25: A large-scale, open-domain multilingual sign language parallel corpus. arXiv preprint arXiv:2407.11144.

Uthus, D., Tanzer, G., & Georg, M. (2023). YouTube-ASL: A large-scale, open-domain American Sign Language–English parallel corpus. Advances in Neural Information Processing Systems, 36, 29029–29047.

Wang, J. (2025). Strategic additions in simultaneous interpreting from a signed language into a spoken language. Translation and Interpreting Studies, 20(1), 24–49.

Wong, R., Camgoz, N. C., & Bowden, R. (2025). SignRep: Enhancing Self-Supervised Sign Representations.

Wong, R., Camgöz, N. C., & Bowden, R. (2024). SIGN2GPT: Leveraging large language models for gloss-free sign language translation. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024).

Yin, A., Zhao, Z., Liu, J., Jin, W., Zhang, M., Zeng, X., & He, X. (2021). SimulSLT: End-to-end simultaneous sign language translation. In Proceedings of the 29th ACM International Conference on Multimedia.

Zhang, B., Müller, M., & Sennrich, R. (2023). SLTUNET: A simple unified model for sign language translation. In Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023).

Zhang, Dongbin, et al. (2025). Guava: Generalizable upper body 3d gaussian avatar. Proceedings of the IEEE/CVF International Conference on Computer Vision.

Zheng, J., Chen, Y., Wu, C., Shi, X., & Kamal, S. M. (2021). Enhancing neural sign language translation by highlighting the facial expression information. Neurocomputing, 464.

Zhou, B., Chen, Z., Clapés, A., Wan, J., Liang, Y., Escalera, S., Lei, Z., & Zhang, D. (2023). Gloss-free sign language translation: Improving from visual-language pretraining. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Zhou, H., Zhou, W., Qi, W., Pu, J., & Li, H. (2021a). Improving sign language translation with monolingual data by sign back-translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Zhou, H., Zhou, W., Zhou, Y., & Li, H. (2021b). Spatial-temporal multi-cue network for sign language recognition and translation. IEEE Transactions on Multimedia.

Zuo, R., Potamias, R. A., Ververas, E., Deng, J., & Zafeiriou, S. (2025). Signs as tokens: A retrieval-enhanced multilingual sign language generator. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

7. Language Resource References

Crasborn, O., & Zwitserlood, I. (2008). The Corpus NGT: An online corpus for professionals and laymen. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood, & E. Thoutenhoofd (Eds.), Construction and exploitation of sign language corpora: 3rd Workshop on the Representation and Processing of Sign Languages (pp. 44–49). ELDA, Paris.

Konrad, R., Hanke, T., Langer, G., Blanck, D., Bleicken, J., Hofmann, I., Jeziorski, O., König, L., König, S., Nishio, R., Regen, A., Salden, U., Wagner, S., Worseck, S., Böse, O., Jahn, E., & Schulder, M. (2020). MEINE DGS — annotiert: Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release. Universität Hamburg. https://doi.org/10.25592/dgs.corpus-3.0

Schembri, A., Fenlon, J., Rentelis, R., & Cormier, K. (2014). British Sign Language Corpus Project: A corpus of digital video data and annotations of British Sign Language 2008-2014 (Second Edition). University College London. http://www.bslcorpusproject.org