SignBind-LLM: Multi-Stage Modality Fusion for Sign Language Translation

Marshall Thomas, Edward Fish, and Richard Bowden

University of Surrey, Guildford, GU2 7XH, United Kingdom
marshall.thomas@surrey.ac.uk
https://www.surrey.ac.uk/centre-vision-speech-signal-processing

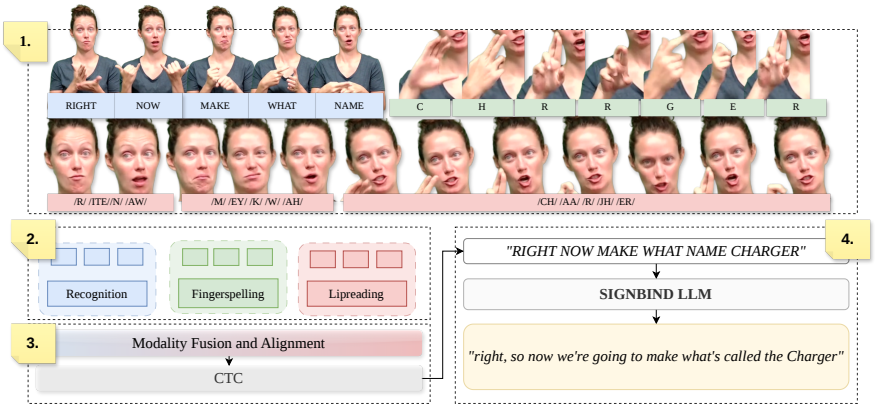

Fig. 1: SignBind-LLM decomposes sign language translation into specialized sub-tasks. We train dedicated experts for continuous sign recognition, fingerspelling detection, and lipreading via CTC on automatically generated pseudo-glosses, fuse them with learned temporal alignment, and decode with a pre-trained LLM. This modular approach substantially outperforms end-to-end alternatives while requiring orders of magnitude less compute.

Abstract. Current sign language translation (SLT) systems attempt to learn all aspects of signing—manual gestures, high-speed fingerspelling, and asynchronous non-manual facial cues—within a single end-to-end network. Learning multiple tasks without detailed supervision leads to poor recognition of fingerspelled proper nouns and technical terms, and leaves rich disambiguating information from lip movements largely unexploited. We introduce **SignBind-LLM**, a modular framework that addresses these limitations through three dedicated expert streams: one for continuous signing, one for fingerspelling, and one for lipreading. Each expert is pre-trained independently using CTC on approximately two million automatically generated pseudo-gloss sequences, removing the need for manual gloss annotation. A lightweight transformer with learned temporal alignment fuses the expert outputs, and a pre-trained language model translates the resulting pseudo-gloss sequences into fluent spoken English. At matched decoder scale (250M parameters), our architecture already surpasses all prior methods, confirming that the gains are archi-

tectural rather than a consequence of scaling the language model. Scaling to a larger decoder sets a new state-of-the-art across How2Sign: 23.1, BOBSL: 7.0, and ChicagoFSWild+: 73.6%, while requiring significantly lower training cost than prior approaches.

Keywords: Sign Language Translation · Fingerspelling · Lipreading · Multi-Stream Architecture · LLMs · Multimodal Fusion

1 Introduction

Sign languages are fully-fledged natural languages used by millions of Deaf people worldwide. They rely simultaneously on manual gestures—handshapes, movements, and spatial locations—and non-manual signals such as facial expressions, mouthings, and mouth gestures[1] Crucially, these channels operate asynchronously: a signer's lip movements do not temporally coincide with their manual articulations, and fingerspelling proceeds at a fundamentally different rate than continuous signing. The field has progressed from isolated sign recognition through continuous recognition to gloss-free sign language translation (SLT), driven by large video–text datasets [9, 19, 50] and the integration of large language models [28]. While gloss-free SLT with LLM's have shown improvement over gloss-based CTC approaches, they exhibit two major limitations that limit their real world application.

Fingerspelling. Fingerspelling constitutes 12–35% of ASL [51] and is the primary mechanism for proper nouns, technical terms, and loanwords—precisely the content that often matters most in a translation. Current gloss-free SLT systems either treat fingerspelling as a subordinate task within the main translation pipeline [27,30] or ignore it entirely [12,26]. The consequence is systematic mistranslation of names, places, and technical vocabulary.

Non-manual cues. Mouthings carry crucial disambiguating information—akin to visemes in speech [10, 34]—and can account for up to 40% of the linguistic information in natural sign production [2]. Yet existing SLT models rarely utilise these signals. The central difficulty is not that the information is absent from the video, but that it is *temporally misaligned* with the manual signs: mouthings may precede or lag behind the corresponding hand movements, and this offset varies continuously. Naïve feature concatenation at the frame level therefore fails to capture the correspondence.

These two challenges share a common cause. In contrast to automatic speech recognition [29] and visual speech recognition [49], where massive pre-training on dedicated modalities is standard practice, current SLT systems force a single network to simultaneously learn continuous signing, high-speed fingerspelling, and asynchronous non-manual cues. This requires a single model to simultaneously learn three distinct recognition problems. Furthermore, training large

[1] In sign language linguistics, *mouthing* denotes silent articulation of spoken words accompanying signs, while *mouth gestures* are non-verbal movements integral to the sign itself, conveying grammatical or affective nuance.

LLMs directly on visual features introduces significant computational overhead and complicates scaling to larger models.

To overcome these limitations, we introduce **SignBind-LLM**, a framework built on a simple principle: *decompose, specialise, then fuse*. We make the following contributions:

1. **Multi-stream expert architecture.** We introduce three dedicated processing streams—for continuous signing, fingerspelling, and lipreading—each pre-trained independently via CTC loss to generate independent text . A dynamic router with straight-through Gumbel-Softmax assigns video segments to the appropriate manual expert during both training and inference.
2. **Automatic pseudo-gloss generation.** We generate and release approximately two million synthetic pseudo-gloss sequences from subtitle text using GPT-4o [38], providing weak CTC supervision without any manual gloss annotation.
3. **Temporal-aware fusion.** A cross-attention mechanism with learned projection matrices dynamically fuses the manual and non-manual streams, handling variable temporal asynchrony without fixed alignment assumptions. We show empirically that this learned alignment substantially outperforms any fixed temporal offset.
4. **LLM decoding.** A pre-trained encoder–decoder LM translates pseudo-gloss sequences into fluent English, trained independently on text pairs and therefore fully decoupled from the visual pipeline.

We evaluate on How2Sign [9], ChicagoFSWild+ [3], and BOBSL [1], requiring only 259 GPU-hours of training—124× less than SSVP-SLT [39]. Critically, at matched decoder scale (250M parameters), SignBind-LLM already outperforms all prior methods (16.5 vs. 15.5 B-4 on How2Sign), demonstrating the gains are architectural. Scaling to a larger decoder yields further improvements (+7.6 B-4 on How2Sign, +4.4 B-4 on BOBSL, +2.1 pp letter accuracy on ChicagoFSWild+), though this comparison is not parameter-matched with prior methods.

2 Related Work

2.1 Sign Language Translation

Early SLT methods relied on manually annotated glosses as an intermediate representation. Gloss-based pipelines [5, 55, 56] use CTC-based continuous sign language recognition (CSLR) [17, 24, 32] to predict gloss sequences, which are then translated to spoken language via sequence-to-sequence models [4]. While effective on controlled benchmarks, these approaches inherit errors from the gloss recogniser and require costly aligned annotations that are unavailable for most sign languages.

Gloss-free methods bypass this bottleneck by learning direct video-to-text mappings. Large-scale video–text pre-training [28, 48, 58, 59] and contrastive objectives [18] have improved translation fluency considerably. Recent work has

focused on leveraging LLMs for decoding sign features, whether from RGB [50], pose [12], or fused modalities [26, 52]. However, all of these systems attempt to learn signing, fingerspelling, and non-manual cues within a single network, leading to systematic underperformance on fingerspelled content and underuse of facial information.

Several multimodal approaches have fused RGB, pose, and optical flow [12, 22], and fingerspelling-specific methods have explored synthetic augmentation [13], large-scale in-the-wild collections [16, 57], and pose-based models [11]. None of these integrates lipreading, despite the well-documented co-articulation between handshapes and lip movements in natural signing [2] and the persistent errors that arise from handshape ambiguity alone.

2.2 Visual Speech Recognition for Signing

Visual speech recognition (VSR) has matured rapidly through end-to-end architectures such as LipNet [40], transformer-based models like Lipformer [53], and large-scale audiovisual pre-training with AV-HuBERT [43,49]. More recent work has integrated LLM-based encoders to improve contextual transcription [54], achieving strong performance even under challenging conditions. In the context of sign language, LLMs have also been used for fingerspelling detection [41, 47] by aligning hand gestures with textual representations.

Despite the maturity of VSR and the availability of large sign language datasets, no prior work has combined lipreading with both continuous sign recognition and fingerspelling detection in a unified framework for SLT—let alone one that explicitly accounts for the temporal misalignment between these streams. Our work fills this gap.

3 Method

We propose **SignBind-LLM**, a sign language translation framework that decomposes the complex translation task into specialized sub-tasks before fusion. The key observation is that sign languages employ distinct communication channels—continuous signing, fingerspelling, and mouthings—that operate at different temporal scales and are asynchronous with respect to one another. Attempting to learn all of these within a single encoder conflates fundamentally different recognition problems. Instead, we assign each channel to a dedicated expert, fuse their outputs with learned temporal alignment, and decode with a pre-trained language model. The pipeline consists of four stages: (1) target generation, (2) modality-specific pre-training, (3) multi-modal fusion, and (4) language model decoding. Figure 2 illustrates the complete architecture.

3.1 Problem Formulation

Given a sign language video $\mathbf{X} = \{x_1, \ldots, x_T\}$ with $x_t \in \mathbb{R}^{H \times W \times 3}$, our objective is to produce an English translation $\mathbf{S} = \{s_1, \ldots, s_L\}$. We decompose this into

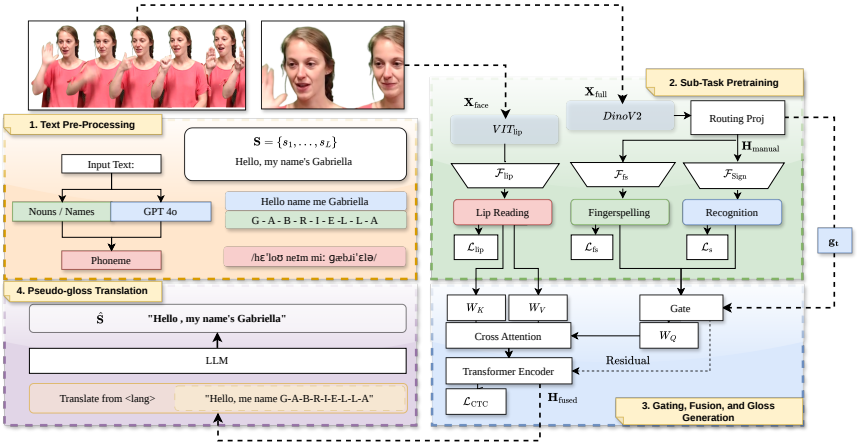

Fig. 2: Overview of SignBind-LLM. Stage 1: Pseudo-glosses and phonemes are generated automatically from subtitles; full-frame and face-cropped video streams are extracted in parallel. Stage 2: specialized expert networks for signing, fingerspelling, and lipreading are pre-trained independently via CTC; a dynamic segment classifier routes video frames to the appropriate manual expert. Stage 3: the expert outputs are gated by the router and fused via cross-attention with learned temporal alignment. Stage 4: a fine-tuned encoder–decoder LLM translates the fused pseudo-gloss sequence into spoken English.

four stages: target-generation, parallel expert encoding, fusion, and language model decoding:

$$\mathbf{S} = \mathrm{LLM}\Big(\mathrm{Fuse}\big(\mathcal{E}_{\mathrm{sign}}(\mathbf{X}),\ \mathcal{E}_{\mathrm{fs}}(\mathbf{X}),\ \mathcal{E}_{\mathrm{lip}}(\mathbf{X})\big)\Big) \tag{1}$$

where $\mathcal{E}_{\mathrm{sign}}$, $\mathcal{E}_{\mathrm{fs}}$, and $\mathcal{E}_{\mathrm{lip}}$ denote the expert encoders for continuous signing, fingerspelling, and lipreading respectively, Fuse reconciles their asynchronous outputs into a unified pseudo-gloss representation $\hat{\mathbf{G}}$, and the LLM translates $\hat{\mathbf{G}}$ into the target sentence.

3.2 Stage 1: Target Generation

A significant barrier to multi-stream modelling is the lack of parallel, granular annotations. We overcome this by automatically generating three complementary training targets from existing English subtitles $\mathbf{S} = \{s_1, \ldots, s_L\}$.

Pseudo-gloss generation. We employ GPT-4o [38] to transform English sentences into sign-language-ordered pseudo-glosses:

$$\mathbf{G} = \mathrm{LLM}_{\mathrm{gloss}}(\mathbf{S}) = \{G_1, \ldots, G_M\} \tag{2}$$

where $M \leq L$ reflects the removal of semantically redundant elements such as articles and the reordering toward subject–object–verb structure. We generate approximately two million such sequences across all training corpora.

Phoneme extraction. To supervise the lipreading module, we extract phoneme sequences from the pseudo-glosses $\mathbf{G}$ using an off-the-shelf phonemiser: $\mathbf{P} = \text{Phonemize}(\mathbf{G}) = \{p_1, \ldots, p_N\}$, where each p_i represents an English phoneme.
Fingerspelling labels. We use the ChicagoFSWild+ [3] training set, decomposing word labels into character sequences $\mathbf{F} = \{f_1, \ldots, f_n\}$ where each f_i is a letter of the alphabet. We additionally identify potential fingerspelling segments in the pseudo-glosses via rule-based detection of proper nouns and technical terms, which are typically fingerspelled in both ASL and BSL.

3.3 Stage 2: Modality-Specific Encoders

This stage defines the expert networks that learn to recognise each communication channel independently.
Video pre-processing. We process the input video into two streams using MediaPipe [31]: full frames $\mathbf{X}_{\text{full}}$ for manual sign recognition and tightly-cropped face regions $\mathbf{X}_{\text{face}}$ for lipreading. Both streams are normalised to 224×224 resolution.
Shared backbone and manual experts. The core of our visual encoder is a DINOv2 [33] backbone that processes $\mathbf{X}_{\text{full}}$ in 16-frame segments, producing shared features $\mathbf{H}_{\text{manual}} \in \mathbb{R}^{B \times T \times D}$. These features serve as input to three parallel components: a dynamic segment router, a continuous sign recognition head, and a fingerspelling detection head.
Dynamic segment routing. Applying all expert heads to all frames is both inefficient and a source of noise—for instance, the fingerspelling head will produce spurious outputs on frames containing continuous signing. We therefore learn a lightweight MLP classifier that routes each temporal segment to one of $\{\text{Sign}, \text{Fingerspelling}, \text{Rest}\}$ and labelled as 'routing proj' in Fig. 2. We first compute a segment-level representation $\bar{\mathbf{h}}_t \in \mathbb{R}^D$ by averaging the DINOv2 features over the frames in segment t, then compute class logits:

$$\bar{\mathbf{h}}_t = \frac{1}{|\mathcal{S}_t|} \sum_{i \in \mathcal{S}_t} \mathbf{H}_{\text{manual},i}\,, \qquad \hat{\mathbf{c}}_t = \mathbf{W}_{\text{cls}}\, \bar{\mathbf{h}}_t + \mathbf{b}_{\text{cls}} \;\in\; \mathbb{R}^3 \tag{3}$$

where $\mathcal{S}_t$ denotes the set of frame indices belonging to segment t.

To obtain discrete routing decisions while maintaining gradient flow during training, we use the straight-through Gumbel-Softmax estimator [20]. The classifier achieves 85% accuracy on this three-way task. Segment labels are assigned as follows: *Fingerspelling* segments are sourced from ChicagoFSWild+ annotations; *Sign* segments are those falling within a non-empty pseudo-gloss region; and *Rest* segments correspond to background or transition frames. In the forward pass, a hard one-hot vector is produced via argmax; in the backward pass, gradients are computed through the continuous softmax relaxation:

$$\mathbf{r}_t = \text{ST-GumbelSoftmax}(\hat{\mathbf{c}}_t,\, \tau) \;\in\; \{0, 1\}^3 \tag{4}$$

where $\mathbf{r}_t = [r_t^{\text{sign}},\ r_t^{\text{fs}},\ r_t^{\text{rest}}]$ and τ is annealed from 1.0 to 0.1 over the course of training. At inference, argmax is applied directly. The routing vector $\mathbf{r}_t$ is passed to the fusion stage (Stage 3) to gate the expert outputs.

Continuous sign recognition. In parallel with the router, the continuous signing expert processes the DINOv2 features. We add learned positional embeddings $\mathbf{PE} \in \mathbb{R}^{T \times D}$ to obtain expert features $\mathbf{F}_{\text{sign}} \in \mathbb{R}^{B \times T \times D}$, then project through a CTC head to the sign vocabulary $\mathcal{V}_{\text{sign}}$:

$$\mathbf{F}_{\text{sign}} = \mathbf{H}_{\text{manual}} + \mathbf{PE}\,, \qquad \hat{\mathbf{L}}_{\text{sign}} = \mathbf{F}_{\text{sign}}\,\mathbf{W}_{\text{sign}} + \mathbf{b}_{\text{sign}} \;\in\; \mathbb{R}^{B \times T \times |\mathcal{V}_{\text{sign}}|} \tag{5}$$

Fingerspelling detection. The fingerspelling expert shares the same position-encoded features but uses an independent CTC head projecting to the 26 characters of the English alphabet:

$$\mathbf{F}_{\text{fs}} = \mathbf{H}_{\text{manual}} + \mathbf{PE}\,, \qquad \hat{\mathbf{L}}_{\text{fs}} = \mathbf{F}_{\text{fs}}\,\mathbf{W}_{\text{fs}} + \mathbf{b}_{\text{fs}} \;\in\; \mathbb{R}^{B \times T \times 26} \tag{6}$$

Lipreading module. The lipreading expert processes the face-cropped stream $\mathbf{X}_{\text{face}}$ independently. Note that this stream captures *mouthings*—speech-derived lip movements that co-occur with signs—rather than the full range of sign-language mouth gestures (which include non-speech movements integral to grammatical marking). This is a deliberate design choice: mouthings carry phonemic information that can be directly supervised with phoneme targets, whereas mouth gestures would require separate annotation. We employ a masked ViT (ViT_{lip}) with a 50% masking ratio to force robust representation learning as shown to be effective in [49]. A 1D convolution then performs temporal downsampling to produce expert features $\mathbf{F}_{\text{lip}}$, and a final linear CTC head projects to the phoneme vocabulary $\mathcal{P}$:

$$\begin{aligned} \mathbf{H}_{\text{lip}} &= \text{ViT}_{\text{lip}}(\text{Mask}(\mathbf{X}_{\text{face}},\, 0.5)) \;\in\; \mathbb{R}^{B \times T \times D_{\text{lip}}} \\ \mathbf{F}_{\text{lip}} &= \text{Conv1D}(\mathbf{H}_{\text{lip}}) \;\in\; \mathbb{R}^{B \times T' \times D'} \\ \hat{\mathbf{L}}_{\text{lip}} &= \mathbf{F}_{\text{lip}}\,\mathbf{W}_{\text{lip}} + \mathbf{b}_{\text{lip}} \;\in\; \mathbb{R}^{B \times T' \times |\mathcal{P}|} \end{aligned} \tag{7}$$

where $T' < T$ due to the convolutional stride, D_{lip} is the ViT hidden dimension, and D' is the dimension of the temporally downsampled features.

All three CTC heads ($\hat{\mathbf{L}}_{\text{sign}}, \hat{\mathbf{L}}_{\text{fs}}, \hat{\mathbf{L}}_{\text{lip}}$) and the segment classifier are trained independently using CTC loss on the targets generated in Stage 1. During fusion (Stage 3), the CTC heads are discarded and the penultimate expert features ($\mathbf{F}_{\text{sign}}, \mathbf{F}_{\text{fs}}, \mathbf{F}_{\text{lip}}$) are used directly.

3.4 Stage 3: Temporal-Aware Multi-Modal Fusion

This stage reconciles the asynchronous expert outputs into a unified representation. The CTC heads used during pre-training are discarded, and the penultimate expert embeddings—which retain richer representational structure than the low-dimensional logits—are projected to a common dimension d via learned linear maps $\phi_{\text{sign}} : \mathbb{R}^{D} \to \mathbb{R}^{d}$, $\phi_{\text{fs}} : \mathbb{R}^{D} \to \mathbb{R}^{d}$, $\phi_{\text{lip}} : \mathbb{R}^{D'} \to \mathbb{R}^{d}$. The lipreading features are temporally upsampled from T' to T via linear interpolation:

$$\mathbf{E}_{\text{sign}} = \phi_{\text{sign}}(\mathbf{F}_{\text{sign}}), \;\; \mathbf{E}_{\text{fs}} = \phi_{\text{fs}}(\mathbf{F}_{\text{fs}}), \;\; \mathbf{E}_{\text{lip}} = \phi_{\text{lip}}(\text{Upsample}(\mathbf{F}_{\text{lip}},\, T)) \tag{8}$$

where $\mathbf{E}_{\text{sign}}, \mathbf{E}_{\text{fs}}, \mathbf{E}_{\text{lip}} \in \mathbb{R}^{B\times T\times d}$.

Gated manual aggregation. The routing vector $\mathbf{r}_t$ from Stage 2 selects between expert outputs at each time step, producing a unified manual representation:

$$\mathbf{M}_t = r_t^{\text{sign}} \cdot \mathbf{E}_{\text{sign},t} \;+\; r_t^{\text{fs}} \cdot \mathbf{E}_{\text{fs},t} \;+\; r_t^{\text{rest}} \cdot \mathbf{e}_{\text{null}} \tag{9}$$

where $\mathbf{e}_{\text{null}} \in \mathbb{R}^d$ is a learned null embedding representing rest periods. Since $\mathbf{r}_t$ is one-hot in the forward pass, exactly one branch is active at each time step; gradients still flow to all branches through the Gumbel-Softmax relaxation during this training stage.

Cross-attention fusion. The relative importance of manual signs versus mouthings is strongly context-dependent. During continuous signing, the manual stream carries the primary semantic load; during fingerspelling, where individual handshapes are visually ambiguous, mouthings become the dominant disambiguating signal. We propose a two-level fusion design: first, a *gated manual aggregation* (described above) that routes between manual experts—itself a contribution that substantially outperforms naïve concatenation (see Table 7)—and second, a *cross-attention* mechanism that integrates the lipreading stream. We model this dynamic balance via multi-head cross-attention, where the choice of query and key/value sources encodes a specific inductive bias: each manual-stream position *queries* the lipreading stream to retrieve the most relevant non-manual context, rather than the reverse. This reflects the linguistic structure of signing, where manual articulations form the primary lexical channel and mouthings provide supporting disambiguation [37].

Concretely, we compute N_h parallel attention heads. For each head i, we project the gated manual representation and the lipreading features into queries, keys, and values of dimension $d_h = d/N_h$:

$$\mathbf{Q}^{(i)} = \mathbf{M}\mathbf{W}_Q^{(i)}, \quad \mathbf{K}^{(i)} = \mathbf{E}_{\text{lip}}\mathbf{W}_K^{(i)}, \quad \mathbf{V}^{(i)} = \mathbf{E}_{\text{lip}}\mathbf{W}_V^{(i)} \tag{10}$$

where $\mathbf{W}_Q^{(i)}, \mathbf{W}_K^{(i)}, \mathbf{W}_V^{(i)} \in \mathbb{R}^{d\times d_h}$. The attention weight matrix $\mathbf{A}^{(i)} \in \mathbb{R}^{T\times T}$ is then:

$$\mathbf{A}^{(i)} = \text{Softmax}\left(\frac{\mathbf{Q}^{(i)}{\mathbf{K}^{(i)}}^\top}{\sqrt{d_h}}\right) \tag{11}$$

Entry $A_{t,t'}^{(i)}$ measures how strongly manual-stream position t attends to lipreading position t'. Crucially, no constraint is placed on the temporal relationship between t and t': the model is free to learn that, for instance, the mouthing corresponding to a sign at frame t actually occurs several frames earlier at $t' < t$. This is the mechanism by which the architecture handles the variable asynchrony between channels without requiring a fixed temporal offset.

The N_h heads are concatenated and projected back to dimension d via $\mathbf{W}_O \in \mathbb{R}^{d\times d}$, then combined with a residual connection, layer normalisation (LN), and a position-wise feed-forward network:

$$\begin{aligned} \mathbf{H}' &= \mathbf{M} + \mathbf{W}_O\,[\text{head}_1; \ldots; \text{head}_{N_h}] \\ \mathbf{H}_{\text{fused}} &= \text{LN}(\mathbf{H}') + \text{FFN}(\text{LN}(\mathbf{H}')) \end{aligned} \tag{12}$$

where $\text{head}_i = \mathbf{A}^{(i)}\mathbf{V}^{(i)}$. Different heads can specialise to different temporal offsets or linguistic functions—for example, one head may learn to align mouthings with the onset of the corresponding sign, while another captures co-articulatory overlap between consecutive signs.

Temporal modelling. The fused features $\mathbf{H}_{\text{fused}}$ are passed through a shallow transformer encoder to model long-range temporal dependencies and to resolve any residual misalignment between the expert streams:

$$\mathbf{Z} = \text{TransformerEncoder}(\mathbf{H}_{\text{fused}}) \in \mathbb{R}^{B\times T\times d} \tag{13}$$

The entire fusion module is trained via CTC loss to produce the correct pseudo-gloss sequence:

$$\hat{\mathbf{G}} = \text{CTC-Decode}(\text{Softmax}(\mathbf{W}_{\text{out}}\mathbf{Z} + \mathbf{b}_{\text{out}})) \tag{14}$$

3.5 Stage 4: Language Model Decoding

The fused representation $\hat{\mathbf{G}}$ is a sequence of pseudo-glosses that follows sign-language word order and lacks function words. It is not a well-formed English sentence. The final stage employs a fine-tuned LLM (Flan-T5-XL [8]) to translate this intermediate representation into a spoken English caption: $\hat{\mathbf{S}} = \text{LLM}_\theta(\hat{\mathbf{G}})$.

This model is trained *independently* of the visual pipeline on a large corpus of (pseudo-gloss, English sentence) pairs $(\mathbf{G}, \mathbf{S})$ generated in Stage 1, using standard cross-entropy loss. The decoupled design has an important practical benefit: because the LLM never sees visual features during training, it learns a robust linguistic mapping that generalises to the noisy, imperfect pseudo-gloss outputs produced by the visual model at inference time. This decoupling also simplifies training and enables flexible deployment with different LLM sizes. It also enables researchers to simply update any component of the architecture without needing to retrain the LLM which has significant benefits in production.

3.6 Training Strategy

Training proceeds in strictly staged phases to ensure stable convergence and component specialisation.

Phase 1 (Pre-training): (i) Each expert head and the segment classifier are trained independently on large-scale data (YouTube-ASL [50] for sign recognition, ChicagoFSWild+ [3] for fingerspelling, and BOBSL [1] for BSL) using their respective CTC losses. (ii) Expert weights are then *frozen* and the fusion module is trained on their outputs, again via CTC. (iii) Separately, the LLM is pre-trained on all available $(\mathbf{G}, \mathbf{S})$ text pairs.

Phase 2 (Fine-tuning): The same staged sequence is repeated on the target evaluation dataset (*e.g.* How2Sign): experts are fine-tuned first, then frozen, then the fusion module is fine-tuned on their updated outputs. The LLM remains frozen throughout Phase 2, having already learned a robust pseudo-gloss-to-English mapping. This decoupled curriculum ensures each component becomes

a strong specialist at its sub-task before its outputs are consumed by downstream modules, and avoids the expensive joint optimisation over all parameters that end-to-end methods require.

4 Experiments

4.1 Datasets

We evaluate on two sign languages (ASL and BSL) across three benchmarks.
YouTube-ASL [50]. A large-scale ASL corpus of 60K videos (∼1,000 h) from over 2,500 signers across diverse backgrounds, lighting conditions, and signing styles. The content is not interpreter-mediated, preserving the diversity of natural signing. We use YouTube-ASL for visual–language pre-training.
How2Sign [9]. A continuous ASL dataset derived from instructional videos, totalling over 80 h across green-screen and studio settings. Eleven signers produced 35,191 sentence-level clips (avg. 162 frames, 17 words) with a vocabulary exceeding 16,000 English words. We fine-tune and evaluate on the standard How2Sign train/test splits.
ChicagoFSWild+ [3]. An in-the-wild fingerspelling dataset containing 55,232 sequences from 260 signers, with varied environments and frequent occlusions. We use the standard train/test split for both pre-training the fingerspelling head and evaluation.
BOBSL [1]. A BSL dataset sourced from BBC broadcast footage, comprising 1,467 h across 1,962 episodes interpreted by 39 signers, with ∼1.2M weakly-aligned English subtitle sentences. We use the improved alignments generated from the method described in [21] and the train/test splits as defined in [1].

4.2 Metrics

We report **BLEU-1** (B-1) and **BLEU-4** (B-4) [36] for translation quality, **ROUGE-L** (R-L) [14] for recall-oriented overlap, and **Letter Accuracy** (L-Acc.) [15] for fingerspelling recognition. Higher is better throughout.

4.3 Implementation Details

All experiments use Flan-T5-XL (3.0B) as the LLM decoder unless otherwise stated. The DINOv2-Base backbone processes 16-frame segments at 224×224 resolution. The fusion transformer has 4 layers with 8 attention heads and dimension d=512. The Gumbel-Softmax temperature τ is annealed from 1.0 to 0.1 over 50K steps. We use AdamW with cosine learning rate scheduling. Total training cost is 259 GPU-hours on a single A100 GPU. Code, all generated pseudo-gloss data, and all pre-trained models will be released on publication.

Table 1: Sign language translation on How2Sign. SignBind-LLM$_{\text{base}}$ (250M) already outperforms all prior methods—including those with 582M decoders and additional pose input. When using the same decoder size as prior work (Flan-T5-Base, 250M), our method already improves BLEU-4 from 15.5 to 16.5—demonstrating that the multi-stream architecture, not merely a larger LM, drives the improvement. Scaling to Flan-T5-XL (3.0B) sets a new SOTA with +7.6 B-4 over SSVP-SLT. Underline: previous best. **Blue** : best overall.

Method	LM Decoder	#P	Input		How2Sign Test			Hrs
			P	R	B-1	B-4	R-L	
GloFE-VN [28]	–	–	✓		14.9	2.2	12.6	–
MSLU [59]	mT5-B	582M	✓	✓	20.1	2.4	17.2	–
SLT-IV [48]	Trans.	≈30M		✓	34.0	8.0	–	–
C^{2}RL [6]	mBART	680M		✓	29.1	9.4	27.0	–
FLa-LLM [7]	mBART	680M		✓	29.8	9.7	–	–
YouTube-ASL [50]	T5-B	250M	✓		37.8	12.4	–	–
SignMusk. [18]	T5-B	250M		✓	41.5	14.3	–	880
Uni-Sign [26]	mT5-B	582M	✓	✓	40.2	14.9	36.0	–
Geo-Sign [12]	mT5-B	582M	✓		40.8	15.1	35.4	–
SSVP-SLT [39]	T5-B	250M		✓	43.2	15.5	38.4	32k
Ours$_{\text{base}}$	FT5-B	250M		✓	45.0	16.5	31.5	
Ours$_{\text{large}}$	FT5-L	780M		✓	46.7	18.6	36.9	
Ours$_{\text{xl}}$	**FT5-XL**	**3.0B**		✓	**50.2**	**23.1**	**42.7**	216/43

#P = LM decoder parameters. P = Pose, R = RGB. FT5 = Flan-T5. B-1/B-4 = BLEU-1/4. R-L = ROUGE-L. Hrs = Visual Encoder GPU hours/LLM GPU-hours.

5 Results

5.1 Sign Language Translation

How2Sign. Table 1 compares SignBind-LLM against recent methods. At the *base* scale (Flan-T5-Base, 250M), our model already achieves 16.5 B-4, surpassing the previous SOTA (SSVP-SLT, 15.5 B-4) with an identically-sized decoder—demonstrating that the gains arise from the multi-stream architecture, not a larger LM. However, we also show that our model scales gracefully as we increase the LLM size. For example, Flan-T5-XL (3.0B) yields **50.2** B-1, **23.1** B-4, and **42.7** R-L (+7.6 B-4 over SSVP-SLT), with significantly less training cost. Our method uses 216 A100 GPU hours to train the visual encoders and just 43 hours of fine-tuning on the LLM. The advantage of our decomposed approach is that the visual encoders can be trained just once and then are interchangeable with other LLM sizes (as shown in the table and supplementary).

BOBSL. Table 2 presents results on BOBSL. SignBind-LLM achieves **7.0** B-4 and **25.8** R-L, improving B-4 by +4.4 over the previous best [22]. These gains confirm that multi-stream fusion generalises from ASL to BSL without architectural modification.

Table 2: Translation on **BOBSL**. B-1/B-4 = BLEU-1/4, R-L = ROUGE-L.

Method	B-1	B-4	R-L
GFSLT [58]	–	0.6	7.4
Sign2GPT [52]	–	0.9	11.4
Albanie et al. [1]	12.8	1.0	10.2
Sincan et al. [46]	18.8	1.3	8.9
Lost in Trans. [22]	–	2.6	15.6
Ours	**28.5**	**7.0**	**25.8**

Table 3: Fingerspelling on **ChicagoFSWild+** test set. L-Acc = letter accuracy.

Method	L-Acc
FS Recognition [45]	41.2
Iterative Attn. [44]	46.7
TDC-SL [35]	50.0
CtoML [25]	54.9
FS Attention [23]	57.8
FSS-Net [42]	64.4
FS PoseNet [11]	71.1
Ours	**73.6**

5.2 Qualitative Analysis

Table 4 presents representative translations. Examples 1–2 show how noisy pseudo-glosses from individual experts are resolved into coherent English after fusion incorporates lipreading phonemes and the LLM adds function words. Example 3 illustrates a failure mode: the ambiguous fused glosses `TODAY WE WORK LOWER HOUSE` cause the LLM to hallucinate "basement" instead of the reference "lower body"—a limitation of the two-stage design where the LLM conditions only on intermediate text.

5.3 Fingerspelling Recognition

Table 3 reports letter accuracy on ChicagoFSWild+. SignBind-LLM achieves **73.6%**, a 2.1 pp improvement over FS PoseNet [11], despite not being specifically designed as a fingerspelling system. The gains arise from lipreading cues during fusion that disambiguate visually similar handshapes; removing the lipreading stream causes an 18.9 pp drop as discussed in Sec. 6. Precision and recall on ChicagoFSWild+ are **76.2%** and **70.4%** respectively. The benefit of dedicated fingerspelling handling is further confirmed by a part-of-speech analysis (supplementary): proper noun accuracy (PROPN) reaches 0.56 for SignBind-LLM versus 0.40 for C^2RL and 0.28 for Geo-Sign—a 40% relative improvement—directly attributable to fingerspelling recognition of names, locations, and technical terms.

6 Ablation Studies

All ablations use the full SignBind-LLM configuration with Flan-T5-XL on the How2Sign test set.

Component ablation. Table 5 reports performance when each component is individually removed. The fusion transformer and LLM produce the largest drops (-13.5 and -14.3 B-4), confirming that cross-modal reconciliation and linguistic reordering are both essential. Among inputs, lipreading has the strongest

Table 4: Qualitative examples from How2Sign (1, 3) and BOBSL (2). *Fused* = pseudo-gloss output from the fusion stage $\mathbf{H}_{\text{fused}}$; *Ours* = final LLM translation; *Ref* = ground-truth subtitle. Examples 1–2 succeed; Example 3 shows semantic drift. Extended outputs in supplementary.

Example 1 — How2Sign ✓	
F	`GAIN BALANCE ON TOE YOU HAVE CENTER YOU SHOULDERS LOOK UP MOUNTAIN YOU CAME FROM WHERE`
O	So to gain that balance on your toe, all you have to do is center your shoulders by looking to the mountain. This is where you came from.
R	To gain balance on your toe side, all you have to do is center your shoulders by looking up the mountain, back where you came from.
Example 2 — BOBSL ✓	
F	`AN FORTUNATELY CATS BEST KEEP LITTLE PARROT ALIVE TAKE OUT`
O	Unfortunately the cats are best kept outside to keep the little parrot alive.
R	Unfortunately for the cats, the best way to keep this chubby little parrot alive is to take kitty out of the equation.
Example 3 — How2Sign ✗	
F	`TODAY WE WORK LOWER HOUSE`
O	Today we're working on the lower part of the house, the basement.
R	Today we're going to work on stretching and strengthening the lower body.

F = Fused pseudo-glosses, O = Our translation, R = Reference.

Table 5: Component ablation. Fusion (−13.5 B-4) and LLM (−14.3 B-4) cause the largest drops. Among inputs, lipreading is most important (−8.5 B-4). L-Acc = letter accuracy (%).

Variant	L-Acc	B-4
W/o Lipreading	54.7	14.6
W/o Fingerspelling	34.1	19.9
W/o Fusion	50.2	9.6
W/o Router	51.6	15.3
W/o LLM	59.2	8.8
Full Model	**73.6**	**23.1**

Table 6: Temporal alignment between manual and non-manual streams. L-Acc = letter accuracy (%). Learned alignment outperforms all fixed offsets. Negative shifts help slightly (mouthings precede signs).

Variant	L-Acc	B-4
No shift	68.7	18.2
Δ+5 frames	63.2	13.4
Δ−5 frames	71.0	19.1
Δ+10 frames	56.1	8.8
Δ−10 frames	65.3	16.6
Learned	**73.6**	**23.1**

B-4 impact (−8.5), while removing fingerspelling causes the largest L-Acc. drop (−39.5 pp) but a smaller B-4 decrease (−3.2), reflecting its localised role in proper noun recognition.

Stand-alone expert evaluation. Table 8 isolates each stream. Continuous signing alone achieves only 7.3 B-4, which is expected considering the outputs of this stream are glosses and not english sentences. The lipreading branch reaches 66.0% phoneme accuracy, confirming its role as the primary disambiguating signal. That neither stream alone approaches full-model performance highlights the complementarity of the architecture.

Table 7: Fusion and pre-training ablation. Top: fusion strategies. Cross-attention shows additional benefit while without gating the performance reduces significantly (concat+MLP). Bottom: We report zero-shot performance on How2Sign under various training setups. [50]. pt = pre-trained T5, np = from scratch.

Variant	B-4
Fusion strategy	
Concat + MLP	12.4
Gated Fusion	22.3
Gated Fusion + Cross Attn	**23.1**
Zero-shot (YT-ASL → H2S)	
YT-ASL (np) [50]	1.4
YT-ASL (pt) [50]	4.0
Ours	**8.3**
How2Sign only (no YT-ASL)	
Álvarez et al. [9]	2.2
GloFE-VN [28]	2.2
Tarrés et al. [48]	8.0
Ours	**13.7**
Pre-train + Fine-tune	
YT-ASL (pt) [50]	12.4
Ours	**23.1**

Table 8: Stand-alone expert encoder performance. Quality of CTC outputs from the individual expert streams for sign recognition and lipreading. Pho = phoneme accuracy (%).

Stream	Pho	B-4
Sign only	–	7.3
Lipreading only	66.0	–

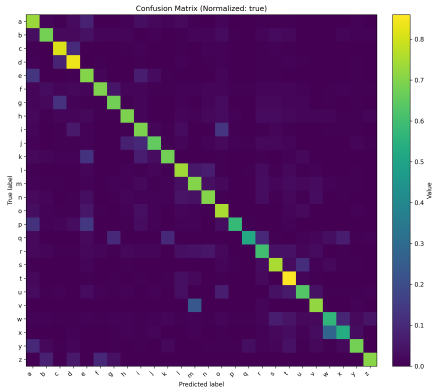

Fig. 3: Fingerspelling confusion matrix on ChicagoFSWild+ (normalised)

Fusion strategy. Table 7 (top) compares three fusion strategies. Concatenation followed by a two-layer MLP achieves only 12.4 B-4—a 10.7-point drop—because it treats all time steps identically and cannot resolve the temporal misalignment between streams. Simply applying gating and then performing concatenation before self-attention (without cross-attention between experts) achieves 23.1 B-4 (-1 B-4).

Pre-training and zero-shot transfer. Table 7 (bottom three sections) compares training strategies against published baselines from [50]. In the zero-shot setting (trained on YT-ASL, evaluated on How2Sign without fine-tuning), our model reaches 8.3 B-4—more than double the best YT-ASL baseline (4.0). Training only on How2Sign yields 13.7 B-4, which already exceeds all prior H2S-only methods including Tarrés et al. (8.0). The full staged pipeline (pre-train on YT-ASL, fine-tune on H2S) achieves 23.1 B-4, nearly doubling the best comparable YT-ASL staged result (12.4). These results demonstrate that our expert architecture extracts more from the same data at every training configuration.

Temporal alignment. Table 6 compares fixed temporal offsets against learned alignment. A small negative shift ($\Delta{-}5$) slightly outperforms naïve frame-aligned fusion, consistent with mouthings preceding manual signs. Large positive shifts degrade performance severely ($\Delta{+}10$: -14.3 B-4), while negative shifts are more benign ($\Delta{-}10$: -6.5 B-4), suggesting greater robustness to mouthings leading

hands than the reverse. The learned alignment mechanism outperforms all fixed offsets by a wide margin, validating the design choice of letting cross-attention learn the correspondence rather than assuming a fixed temporal relationship.

Fingerspelling error analysis. Figure 3 presents the confusion matrix for fingerspelling on ChicagoFSWild+. Interestingly, the dominant error patterns do not always involve visually similar handshapes: `M`/`V` (differing in finger extension). This suggests that performance may be limited by occlusion, resolution, or motion blur rather than the distance between fingers in the image, or that the fingerspelling network is leveraging additional context from the whole body to distinguish between letters. A detailed parts-of-speech analysis demonstrating that our improvements are concentrated on visually grounded lexical categories is provided in the supplementary material.

7 Conclusion

We have presented SignBind-LLM, a modular framework that decomposes sign language translation into dedicated experts for continuous signing, fingerspelling, and lipreading, fused with learned temporal alignment and decoded by a pretrained LLM. This decomposition achieves state-of-the-art results on three benchmarks (**23.1** B-4 on How2Sign, **7.0** B-4 on BOBSL, **73.6%** letter accuracy on ChicagoFSWild+) while requiring only 259 GPU hours, significantly less than existing approaches. Critically, the multi-stream architecture already outperforms prior methods at equal LM scale, confirming the gains are architectural. Ablations reveal complementary contributions from each stream, essential learned temporal alignment, and the advantage of modular training over end-to-end approaches.

Acknowledgements

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

References

1. Albanie, S., Varol, G., Momeni, L., Bull, H., Afouras, T., Chowdhury, H., Fox, N., Woll, B., Cooper, R., McParland, A., Zisserman, A.: BOBSL: BBC-Oxford British Sign Language Dataset (November 2021)

2. Aparicio, M., Peigneux, P., Charlier, B., Balériaux, D., Kavec, M., Leybaert, J.: The neural basis of speech perception through lipreading and manual cues: Evidence from deaf native users of cued speech. Neuropsychologia (2017)
3. B. Shi, A. Martinez Del Rio, J.K.D.B.G.S., Livescu, K.: Fingerspelling recognition in the wild with iterative visual attention. ICCV (October 2019)
4. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
5. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: Joint end-to-end sign language recognition and translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2020)
6. Chen, Z., Zhou, B., Huang, Y., Wan, J., Hu, Y., Shi, H., Liang, Y., Lei, Z., Zhang, D.: C2rl: Content and context representation learning for gloss-free sign language translation and retrieval. IEEE Transactions on Circuits and Systems for Video Technology (2024)
7. Chen, Z., Zhou, B., Li, J., Wan, J., Lei, Z., Jiang, N., Lu, Q., Zhao, G.: Factorized learning assisted with large language model for gloss-free sign language translation. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (May 2024)
8. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E.H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V., Wei, J.: Scaling instruction-finetuned language models. J. Mach. Learn. Res. (2024)
9. Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., Giro-i Nieto, X.: How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
10. Editors, B.: Phoneme. Encyclopedia Britannica (nov 2025), `https://www.britannica.com/topic/phoneme`
11. Fayyazsanavi, P., Nejatishahidin, N., Košecká, J.: Fingerspelling posenet: Enhancing fingerspelling translation with pose-based transformer models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops (January 2024)
12. Fish, E., Bowden, R.: Geo-sign: Hyperbolic contrastive regularisation for geometrically aware sign language translation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
13. Fowley, F., Ventresque, A.: Sign language fingerspelling recognition using synthetic data. In: The 29th Irish Conference on Artificial Intelligence and Cognitive Science 2021, Dublin, Republic of Ireland, December 9-10, 2021 (2021)
14. Ganesan, K.A.: Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. ArXiv (2015)
15. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S.: TIMIT acoustic-phonetic continuous speech corpus. techreport LDC93S1, Linguistic Data Consortium (1993)
16. Georg, M., Tanzer, G., Uboweja, E., Hassan, S., Shengelia, M., Sepah, S., Forbes, S., Starner, T.: Fsboard: Over 3 million characters of asl fingerspelling collected via smartphones. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) (June 2025)

17. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd International Conference on Machine Learning (ICML). ACM (2006)
18. Gueuwou, S., Du, X., Shakhnarovich, G., Livescu, K.: SignMusketeers: An efficient multi-stream approach for sign language translation at scale. In: Findings of the Association for Computational Linguistics: ACL 2025 (Jul 2025)
19. Hanke, T., Schulder, M., Konrad, R., Jahn, E.: Extending the Public DGS Corpus in size and depth. In: Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives (May 2020)
20. Jang, E., Gu, S., Poole, B.: Categorical reparametrization with gumbel-softmax. In: Proceedings International Conference on Learning Representations (ICLR) (2017)
21. Jang, Y., Choi, J., Ahn, J., Chung, J.S.: Deep understanding of sign language for sign to subtitle alignment. IEEE Transactions on Multimedia (2025)
22. Jang, Y., Raajesh, H., Momeni, L., Varol, G., Zisserman, A.: Lost in translation, found in context: Sign language translation with contextual cues. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) (June 2025)
23. Kabade, A.E., Desai, P., C, S., G, S.: American sign language fingerspelling recognition using attention model. In: 2023 IEEE 8th International Conference for Convergence in Technology (I2CT) (2023)
24. Koller, O., Camgoz, N., Ney, H., Bowden, R.: Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. IEEE Transactions on Pattern Analysis and Machine Intelligence (04 2019). https://doi.org/10.1109/TPAMI.2019.2911077
25. Li, L., Jin, T., Cheng, X., Wang, Y., Lin, W., Huang, R., Zhao, Z.: Contrastive token-wise meta-learning for unseen performer visual temporal-aligned translation. In: Findings of the Association for Computational Linguistics: ACL 2023 (Jul 2023)
26. Li, Z., Zhou, W., Zhao, W., Wu, K., Hu, H., Li, H.: Uni-sign: Toward unified sign language understanding at scale. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=0Xt7uT04cQ
27. Liang, H., Huang, C., Xu, Y., Tang, C., Ye, W., Zhang, J., Chen, X., Yu, J., Xu, L.: Llava-slt: Visual language tuning for sign language translation. arXiv preprint arXiv:2412.16524 (2024)
28. Lin, K., Wang, X., Zhu, L., Sun, K., Zhang, B., Yang, Y.: Gloss-free end-to-end sign language translation. arXiv preprint arXiv:2305.12876 (2023)
29. Long, Z., Shen, Y., Fu, C., Gao, H., lijiang Li, Chen, P., Zhang, M., Shao, H., Li, J., Peng, J., Cao, H., Li, K., Ji, R., Sun, X.: VITA-audio: Fast interleaved audio-text token generation for efficient large speech-language model. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
30. Low, J.H., Sincan, O.M., Bowden, R.: Sage: Segment-aware gloss-free encoding for token-efficient sign language translation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops (October 2025)
31. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M., Lee, J., Chang, W.T., Hua, W., Georg, M., Grundmann, M.: Mediapipe: A framework for perceiving and processing reality. In: Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019 (2019)

32. Min, Y., Hao, A., Chai, X., Chen, X.: Visual alignment constraint for continuous sign language recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2021)
33. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
34. Ozel, M.: Viseme cheat sheet (2024), `https://melindaozel.com/viseme-cheat-sheet/`
35. Papadimitriou, K., Potamianos, G.: Multimodal sign language recognition via temporal deformable convolutional sequence learning. In: Interspeech 2020 (2020)
36. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (Jul 2002)
37. Pfau, R., Steinbach, M., Woll, B.: Sign Language: An International Handbook. Handbooks of Linguistics and Communication Science, De Gruyter Mouton, Berlin, Germany (2012)
38. Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training. arXiv (2018)
39. Rust, P., Shi, B., Wang, S., Camgoz, N.C., Maillard, J.: Towards privacy-aware sign language translation at scale. In: Annual Meeting of the Association for Computational Linguistics (2024)
40. S, J., Antony, A.: Lipnet: End-to-end lipreading. SSRN Electronic Journal (01 2024)
41. Shi, B.: Toward american sign language processing in the real world: Data, tasks, and methods. ArXiv **abs/2308.12419** (2023)
42. Shi, B., Brentari, D., Shakhnarovich, G., Livescu, K.: Searching for fingerspelled content in American Sign Language. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (May 2022)
43. Shi, B., Hsu, W.N., Lakhotia, K., Mohamed, A.: Learning audio-visual speech representation by masked multimodal cluster prediction. arXiv preprint arXiv:2201.02184
44. Shi, B., Rio, A.M.D., Keane, J., Brentari, D., Shakhnarovich, G., Livescu, K.: Fingerspelling recognition in the wild with iterative visual attention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
45. Shi, B., Rio, A.M.D., Keane, J., Michaux, J., Brentari, D., Shakhnarovich, G., Livescu, K.: American sign language fingerspelling recognition in the wild. 2018 IEEE Spoken Language Technology Workshop (SLT) (2018)
46. Sincan, O.M., Camgoz, N.C., Bowden, R.: Is context all you need? Scaling Neural Sign Language Translation to Large Domains of Discourse . In: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW) (Oct 2023)
47. Tanzer, G.: Fingerspelling within sign language translation. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (Apr 2025)

48. Tarrés, L., Gállego, G.I., Duarte, A., Torres, J., i Nieto, X.G.: Sign language translation from instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2023)
49. Thomas, M., Fish, E., Bowden, R.: Vallr: Visual asr language model for lip reading. In: ICCV (2025)
50. Uthus, D., Tanzer, G., Georg, M.: Youtube-asl: a large-scale, open-domain american sign language-english parallel corpus. In: Proceedings of the 37th International Conference on Neural Information Processing Systems (2023)
51. Valli, C., Lucas, C., Mulrooney, K.J., Villanueva, M.: Linguistics of American Sign Language: An Introduction. Gallaudet University Press, Washington, D.C., 5th edn. (2011)
52. Wong, R., Camgoz, N.C., Bowden, R.: Sign2GPT: Leveraging large language models for gloss-free sign language translation. In: ICLR (2024)
53. Xue, F., Li, Y., Liu, D., Xie, Y., Wu, L., Hong, R.: Lipformer: Learning to lipread unseen speakers based on visual-landmark transformers. IEEE Transactions on Circuits and Systems for Video Technology (2023)
54. Yeo, J., Han, S., Kim, M., Ro, Y.M.: Where visual speech meets language: VSP-LLM framework for efficient and context-aware visual speech processing. In: Findings of the Association for Computational Linguistics: EMNLP 2024 (Nov 2024)
55. Zhang, B., Müller, M., Sennrich, R.: SLTUNET: A simple unified model for sign language translation. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=EBS4C77p_5S
56. Zhao, R., Zhang, L., Fu, B., Hu, C., Su, J., Chen, Y.: Conditional variational autoencoder for sign language translation with cross-modal alignment. In: Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence (2024)
57. Zhidebayeva, A., Nurmukhanbetova, G., Aldeshov, S., Zhamalova, K., Mamikov, S., Torebay, N.: Real-time sign language fingerspelling recognition system using 2d deep cnn with two-stream feature extraction approach. International Journal of Advanced Computer Science and Applications (2024)
58. Zhou, B., Chen, Z., Clapés, A., Wan, J., Liang, Y., Escalera, S., Lei, Z., Zhang, D.: Gloss-free sign language translation: Improving from visual-language pretraining. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2023)
59. Zhou, W., Zhao, W., Hu, H., Li, Z., Li, H.: Scaling up multimodal pre-training for sign language understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)