SIGNET: Motion-Level Knowledge Transfer for Cross-Language Sign Language Translation

Sobhan Asasi, Ozge Mercanoglu Sincan, and Richard Bowden

CVSSP, University of Surrey, United Kingdom
{s.asasi, o.mercanoglusincan, r.bowden}@surrey.ac.uk
https://cogvis-cvssp.github.io/papers/signet

Abstract. Sign language translation (SLT) remains challenging due to its high spatio-temporal complexity, long sequences, and the need to model multiple articulators without relying on gloss annotations. Existing approaches are typically tailored to individual datasets or languages and struggle to scale, while overlooking the relationships between sign languages that could inform more effective cross-lingual transfer. We present **SIGNET**, a framework that enables motion-level knowledge transfer for cross-language sign language translation. Our key insight is that, although sign languages differ in grammar and lexicon, pretrained models capture motion-level visual patterns that can be reused across datasets and languages. **SIGNET** integrates multiple pretrained sign language backbones through an attention-based, hand-prior aggregation mechanism that guides a gated fusion network in dynamically selecting the most relevant experts. Comprehensive experiments on four benchmarks (How2Sign, Phoenix14T, CSL-Daily, and MeineDGS) demonstrate state-of-the-art translation performance, and **SIGNET** also surpasses prior methods on WLASL for sign language recognition.

Keywords: Sign Language Translation · Cross-lingual Transfer

1 Introduction

Sign Language Translation (SLT) aims to generate spoken or written sentences directly from sign language videos. SLT is particularly challenging because it requires the conversion of continuous video sequences with complex spatio-temporal dynamics into coherent sequences of textual tokens [6, 43]. This inherently multimodal problem with long temporal dependencies requires large amounts of training data and computation, limiting scalability and making it difficult for current systems to generalise beyond individual (small) datasets. This raises a fundamental question: to what extent can motion-level knowledge learned from one sign language be reused to benefit translation in another?

To scale to larger data, recent studies have shifted towards gloss-free SLT methods [6, 17, 43, 50, 53], which remove the need for intermediate gloss[1] annotations, and directly translate from video to text.

[1] Gloss is a written representation of a sign, typically denoted by one or more words from a spoken language.

However, most existing approaches are evaluated primarily on two small-scale benchmarks, Phoenix14T [6] and CSL-Daily [59], which leads to overfitting and ignores the importance of generalisation [4, 17, 28, 50, 53].

Most models rely on pretrained transformer- or CNN-based visual encoders that process individual frames independently, making it difficult to subsample or discard frames without losing crucial linguistic information [17, 20, 28, 37, 50, 57]. Moreover, current methods are generally trained and fine-tuned in isolation for each dataset, discarding previously learned representations when adapting to new sign languages [17,20,28,37,50,53,57]. This fragmented training paradigm overlooks the fact that, while sign languages differ in grammar and lexicon across language families, they share underlying articulatory structure, and backbones pretrained on different languages can capture complementary visual cues. Leveraging such transferable representations from large-scale datasets to low-resource languages motivates our proposed method.

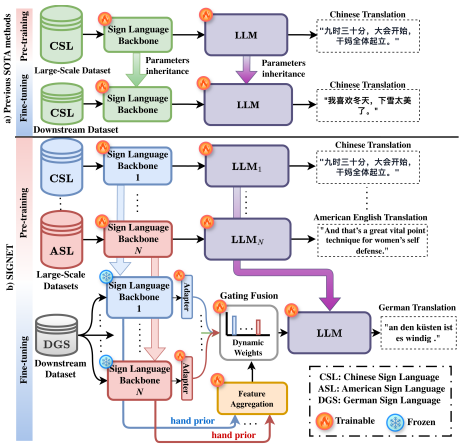

Fig. 1: Comparison between prior SLT methods and our framework. While existing models are language-specific and fail to generalise across other sign languages, **SIGNET** integrates multiple pretrained backbones via adaptive gating.

As shown in Fig. 1, recent state-of-the-art approaches in sign language translation largely depend on extensive pretraining on language-specific datasets [16, 31, 37]. Although such pretraining improves performance on those languages, methods do not explore strategies to transfer the acquired knowledge to other sign languages. As a result, their generalisation is compromised and leads to overfitting to the domains represented in pretraining data.

To address these limitations, we propose **SIGNET** (**SI**gn Lan**G**uage **NET**work), a framework designed to analyse and exploit motion-level knowledge transfer across sign languages by treating pretrained sign backbones as reusable visual experts. Our main contributions are summarised as follows: (***i***) We introduce **SIGNET**, a framework that integrates independently pretrained sign language backbones as frozen visual experts with distinct motion-level expertise, combined through adaptive routing. Unlike standard Mixture-of-Experts (MoE), our experts encode different motion vocabularies from their pretraining languages. (***ii***) We propose hand priors as a domain-specific routing signal, where hand-centric descriptors are aggregated into an input-dependent expert signature that guides gating, introducing articulator-level inductive bias absent in standard MoE routing. (***iii***) We show that frozen cross-lingual experts can be effectively adapted to new sign languages through a combination of low-rank residual updates and contrastive alignment, requiring only 1.5M trainable pa-

rameters in the video encoder while matching methods that demand orders of magnitude more compute. ***(iv)*** Experiments across four SLT benchmarks reveal that language-matched experts dominate when available, while for languages absent from pretraining the framework discovers complementary specialisations across experts. **SIGNET** also surpasses prior methods on WLASL for sign language recognition.

ASL, BSL, DGS, and CSL belong to distinct sign language lineages with limited lexical overlap and also differ in dataset domains (instructional, weather, and daily conversation). This diversity allows us to examine how motion-level priors transfer across both linguistic and domain boundaries.

2 Related Work

Sign Language Translation. SLT aims to convert continuous sign videos into spoken or written language sequences. Existing approaches can be broadly categorised into two paradigms: *gloss-based* and *gloss-free*. Gloss-based methods first predict an intermediate gloss sequence, representing sign-level lexical units, before translating it into text [7, 9, 54, 55, 60, 61]. While this intermediate representation provides linguistic structure, it also introduces a dependency on expensive and labour-intensive gloss annotations, which limits data scalability and model generalisation across languages. To overcome these challenges, recent works have increasingly adopted *gloss-free* frameworks that directly translate visual features into text, a shift largely enabled by the success of transformer architectures [11, 17, 32, 50, 57]. In this direction, several studies [3, 10, 11, 50, 57] employ pretext pretraining and contrastive learning to align visual and linguistic modalities. With the advent of large language models (LLMs), more recent approaches [3, 10, 25, 50] leverage pretrained linguistic priors to improve translation fluency and coherence. Other works [4, 20, 28] incorporate LLMs or Multimodal LLMs directly into pretraining, utilising their reasoning capabilities to extract structured motion and semantic representations. Despite these advances, existing SLT methods remain computationally demanding and dataset-specific, with no mechanism to transfer learned representations across sign languages.

Efficient Sign Language Representations. Sign language data is temporally dense, making conventional downsampling infeasible without information loss. Although transformer- and CNN-based models [4, 10, 11, 20, 28, 50, 57] achieve strong results, their computational demands hinder large-scale or multilingual training. Skeletal representations provide a compact alternative by efficiently encoding signer motion while maintaining linguistic structure. Recent ST-GCN variants [16, 31, 51] have shown promise for sign modelling but remain limited in low-resource scenarios [34, 46, 56]. Our framework builds upon these efficient graph-based encoders as modular backbones and unifies them via adaptive expert fusion, achieving scalable and cross-lingual sign understanding with minimal computational overhead.

Scalability in Sign Language. Acquiring large-scale aligned sign-text data remains challenging due to privacy constraints and limited accessibility. As a result,

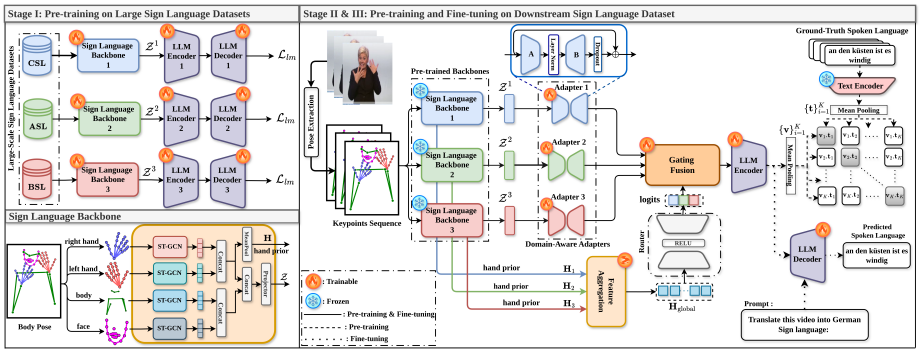

Fig. 2: Overview of **SIGNET**. Stage I pretrains sign language backbones and their LLM encoder-decoders on large-scale datasets (CSL, ASL, and BSL), where each backbone (bottom left) learns part-specific motion cues via ST-GCN modules. Frozen pretrained backbones provide *hand priors* to the feature aggregation module, which computes a global expert descriptor to guide the gating fusion network. Stage II aligns visual and textual spaces via cross-modal pretraining, followed by Stage III fine-tuning. See Fig. 3 for a detailed view of the aggregation and gating modules.

most prior works rely on small and domain-limited datasets such as Phoenix14T and CSL-Daily [4,17,28,50,57], which often lead to overfitting. In contrast, larger and more diverse corpora such as MeineDGS [21] remain underexplored. Recent multilingual datasets [2,31,44] support scalable modelling, and pretraining-fine-tuning pipelines [31,37] have shown improvements, yet their success depends on lexical overlap between corpora and often demands extensive resources (e.g., 64 A100 GPUs in [37]). Rather than assuming uniform transferability across sign languages, our framework leverages the observation that different pretrained backbones encode distinct yet complementary articulation patterns [47], which can be selectively combined to support efficient knowledge transfer across sign language domains.

Expert Integration for Scalability. The mixture-of-experts (MoE) paradigm enables selective activation of relevant experts while maintaining computational efficiency [13,15,19,41,58,62]. Complementary developments in parameter-efficient fine-tuning [22,23,27] and multimodal gating mechanisms [1,12,30] further enhance scalability. Previous models are typically trained in isolation for each dataset and language. Our framework addresses this limitation by treating multiple sign language backbones, each pretrained on a distinct large-scale corpus, as visual experts.

3 Methodology

We first outline the training of multiple sign language backbones on large-scale datasets (Sec. 3.1-3.3), then describe the modules enabling knowledge transfer

to downstream datasets (Sec. 3.4-3.6), and finally present the pretraining and fine-tuning stages for adaptation (Sec. 3.7-3.8).

3.1 Pre-Processing

We extract 2D skeletal keypoints from each video using RTMPose [26] and partition them into four anatomical regions: the body, left hand, right hand, and face, denoted $\{b, lh, rh, f\}$. Each keypoint (x, y) is extended with its detection confidence score to form (x, y, c), where $c \in [0, 1]$. The resulting keypoint sequence for each region r is denoted $\mathcal{P}_r \in \mathbb{R}^{T \times J_r \times 3}$, where T is the number of frames and J_r is the number of joints in that region.

3.2 Sign Language Backbone

The proposed sign language backbone, denoted as $\mathcal{B}$, is designed to jointly capture motion dynamics and articulation cues across different anatomical regions of the signer (as illustrated in Fig. 2, bottom left). It consists of four spatio-temporal graph convolutional network (ST-GCN) modules [51], each responsible for a specific region. For each region, the corresponding keypoint sequence $\mathcal{P}_r$ is processed by its dedicated ST-GCN, $\mathcal{G}_r$, within the backbone $\mathcal{B}$ to extract fine-grained spatio-temporal representations:

$$\mathcal{F}_r = \mathcal{G}_r(\mathcal{P}_r), \quad \mathcal{F}_r \in \mathbb{R}^{T \times C_r}. \tag{1}$$

Formally, the overall backbone can be expressed as $\mathcal{B} = \{\mathcal{G}_b, \mathcal{G}_{lh}, \mathcal{G}_{rh}, \mathcal{G}_f\}$, where each hand ($\mathcal{G}_{rh}$,$\mathcal{G}_{lh}$) operates in parallel to encode distinct motion cues and gestures from the hands, non-manual expressions from the face ($\mathcal{G}_f$), and global body dynamics ($\mathcal{G}_b$). The resulting feature maps $\{\mathcal{F}_r\}$ are later integrated to form a comprehensive signer representation that serves as the visual foundation for translation. See Appendix for more details.

3.3 Stage I: Pretraining on Large-Scale Datasets

As illustrated in Fig. 2, this stage establishes the foundational visual-linguistic experts used throughout the framework. For the i-th large-scale sign language dataset $\mathcal{D}^i = (\mathcal{P}^i, \mathbf{y}^i)$, where $\mathcal{P}^i$ denotes the skeletal keypoint sequences and $\mathbf{y}^i$ the corresponding textual translations, a dedicated backbone $\mathcal{B}^i$ is jointly trained with a multilingual LLM-based encoder-decoder to align spatio-temporal visual dynamics with linguistic semantics in a shared representation space. Each backbone $\mathcal{B}^i$ comprises four region-specific ST-GCN modules $\{\mathcal{G}^i_b, \mathcal{G}^i_{lh}, \mathcal{G}^i_{rh}, \mathcal{G}^i_f\}$ that process the corresponding anatomical inputs, $\{\mathcal{P}^i_b, \mathcal{P}^i_{lh}, \mathcal{P}^i_{rh}, \mathcal{P}^i_f\} \subset \mathcal{P}^i$. For each skeletal keypoint sequence in $\mathcal{P}^i$, the resulting part-level features are concatenated:

$$\mathcal{F}^i_{\text{concat}} = [\mathcal{F}^i_b; \mathcal{F}^i_{lh}; \mathcal{F}^i_{rh}; \mathcal{F}^i_f] \in \mathbb{R}^{T \times C}, \tag{2}$$

where $C = \sum_r C_r$ is the total feature dimension across all regions. A linear projection maps this representation into the language model's embedding space,

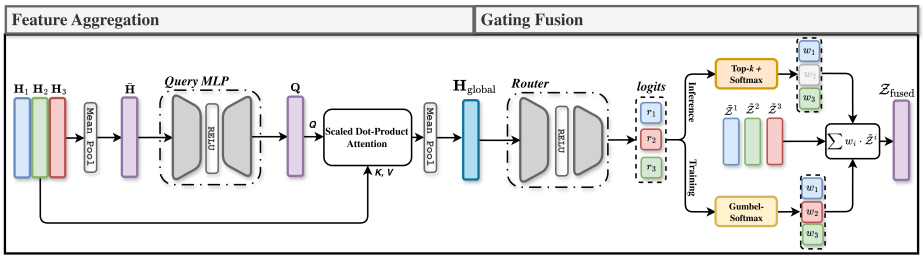

Fig. 3: Feature Aggregation and Gating Fusion modules. The feature aggregation module (left) computes a global expert descriptor $\mathbf{H}_{\text{global}}$ from hand priors via attentive pooling. The gating fusion module (right) routes $\mathbf{H}_{\text{global}}$ through a learned router to dynamically weight expert contributions, using Gumbel-Softmax during training and top-k selection at inference.

yielding $\mathcal{Z}^i \in \mathbb{R}^{T \times E}$, which is fed to the LLM encoder-decoder to predict the target sentence $(y_1^i, \ldots, y_{T'}^i) \in \mathbf{y}^i$ via an autoregressive decoder:

$$\mathcal{L}_{lm} = -\sum_{t=1}^{T'} \log P(y_t \,|\, y_{<t}, \mathcal{Z}^i). \tag{3}$$

where y_t represents the t-th token and $y_{<t}$ represents all previous tokens.

Through this end-to-end training, each backbone $\mathcal{B}^i$ learns to encode complementary spatial and temporal cues, including manual gestures from the hands, non-manual facial expressions, and full-body motion, into linguistically grounded visual representations. These pretrained backbones are subsequently frozen and reused as visual experts.

3.4 Lightweight Expert Adaptation

Given a downstream dataset $\mathcal{D}^* = (\mathcal{P}^*, \mathbf{y}^*)$, where $\mathcal{P}^*$ denotes the skeletal input sequences and $\mathbf{y}^*$ the corresponding spoken-language translations, each input in $\mathcal{P}^*$ is passed through all pretrained backbones $\{\mathcal{B}^1, \mathcal{B}^2, \ldots, \mathcal{B}^N\}$ to obtain expert-level embeddings $\{\mathcal{Z}^1, \mathcal{Z}^2, \ldots, \mathcal{Z}^N\}$ (as shown in Fig. 2) where N is the number of pretrained backbones. While these pretrained experts provide strong visual priors, their projected embeddings $\mathcal{Z}^i$ may still exhibit domain mismatch when transferred to new datasets. To efficiently adapt these representations without updating the full model, we introduce lightweight expert adapters applied directly on top of $\mathcal{Z}^i$. Each adapter refines the frozen expert representations through a low-rank residual transformation $\tilde{\mathcal{Z}}^i = \mathcal{Z}^i + (\text{LN}(\mathcal{Z}^i)\mathbf{A}_i)\mathbf{B}_i$, where $\text{LN}(\cdot)$ denotes layer normalisation, and $\mathbf{A}_i \in \mathbb{R}^{E \times R}$ and $\mathbf{B}_i \in \mathbb{R}^{R \times E}$ form a learnable low-rank decomposition with rank $R \ll E$. Following the LoRA [23] formulation, the residual update $\Delta W = \mathbf{A}_i\mathbf{B}_i$ acts as a compact adaptation path, enabling the model to learn dataset-specific adjustments while keeping the pretrained expert parameters frozen.

3.5 Attention-based Feature Aggregation

To derive an input-dependent global descriptor for adaptive expert routing, for each pretrained backbone $\mathcal{B}^i$, we extract right- and left-hand features $\mathcal{F}^i_{rh}$ and $\mathcal{F}^i_{lh}$ from its corresponding ST-GCN modules, apply temporal mean pooling, and concatenate them to form the *hand prior*:

$$\mathbf{H}_i = [\mathrm{Mean}_T(\mathcal{F}^i_{rh}); \mathrm{Mean}_T(\mathcal{F}^i_{lh})] \in \mathbb{R}^{B\times D}. \tag{4}$$

where B is the batch size and $D = C_{rh} + C_{lh}$.

An attention-based aggregation module (detailed in Alg. 1) then combines the set $[\mathbf{H}_1, \ldots, \mathbf{H}_N]$ into a global descriptor $\mathbf{H}_{\text{global}}$, serving as a soft expert-selection signal that emphasises the most relevant pretrained backbones based on kinematic and semantic cues (see Fig. 3).

3.6 Top-k Gating and Expert Fusion

The global descriptor $\mathbf{H}_{\text{global}}$ is provided to a lightweight router network (as shown in Fig. 2 and Fig. 3), implemented as a two-layer MLP that outputs routing logits across the N pretrained experts:

$$\mathbf{r} = \mathrm{MLP}(\mathbf{H}_{\text{global}}) \in \mathbb{R}^{B\times N}. \tag{5}$$

During training, we employ the Gumbel-Softmax distribution [24, 35] to allow differentiable expert selection while maintaining stochastic exploration. For each expert i, the corresponding routing weight w_i is sampled as:

$$\begin{aligned} w_i &= \frac{\exp((r_i + g_i)/\tau_g)}{\sum_{j=1}^{N} \exp((r_j + g_j)/\tau_g)}, \\ g_i &= -\log(-\log(u_i)), \quad u_i \sim \mathcal{U}(0,1). \end{aligned} \tag{6}$$

where τ_g is the temperature controlling the degree of smoothness, and g_i is the Gumbel noise drawn from a standard Gumbel distribution. As $\tau_g \to 0$, the distribution becomes increasingly discrete, allowing the router to approximate hard expert selection while preserving differentiability during backpropagation. The resulting stochastic weights $\{w_i\}_{i=1}^N$ are used to compute a soft mixture of expert representations:

$$\mathcal{Z}_{\text{fused}} = \sum_{i=1}^{N} w_i \cdot \tilde{\mathcal{Z}}^i. \tag{7}$$

At inference time, we replace the stochastic routing with a deterministic top-k selection strategy. The router selects the k experts with the largest routing logits $\mathbf{r}$, applies a Softmax over their scores, and fuses only those k experts:

$$\mathcal{Z}_{\text{fused}} = \sum_{i\in \text{Top-}k} \frac{\exp(r_i)}{\sum_{j\in \text{Top-}k} \exp(r_j)} \cdot \tilde{\mathcal{Z}}^i. \tag{8}$$

This hybrid training–inference design enables two complementary behaviours: (1) stochastic exploration of expert specialisation during training, and (2) deterministic and interpretable expert routing at test time.

Algorithm 1 Attention-based Feature Aggregation

Require: Hand priors: $\{\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_N\}$, $\mathbf{H}_i \in \mathbb{R}^{B\times D}$

$\mathbf{H} \leftarrow [\mathbf{H}_1; \mathbf{H}_2; \ldots; \mathbf{H}_N]$ ▷ Stack: $\mathbf{H} \in \mathbb{R}^{B\times N\times D}$

$\bar{\mathbf{H}} \leftarrow \text{MeanPool}(\mathbf{H}, \text{dim} = 1)$

$\mathbf{Q} \leftarrow \text{MLP}(\bar{\mathbf{H}})$ ▷ Data-dependent query

$\mathbf{L} \leftarrow (\mathbf{Q}\mathbf{H}^\top)/\sqrt{D}$ ▷ Scaled dot-product attention

$\boldsymbol{\alpha} \leftarrow \text{Softmax}(\mathbf{L}, \text{dim} = -1)$

$\mathbf{H}_{\text{global}} \leftarrow \boldsymbol{\alpha}\mathbf{H}$ ▷ Weighted sum over backbones

$\mathbf{H}_{\text{global}} \leftarrow \text{LayerNorm}(\mathbf{H}_{\text{global}})$

return $\mathbf{H}_{\text{global}}$

3.7 Stage II: Cross-Modal Pretraining

When adapting to a downstream sign language dataset, the training path depends on whether its spoken-language domain was present during large-scale pretraining. If the language has been seen, we directly reuse the corresponding pretrained LLM encoder-decoder and proceed to Stage III. For languages not covered during pretraining, one of the pretrained LLM encoder-decoders in Stage I is randomly selected and further aligned via a cross-modal contrastive pretraining stage, which bridges the gap between visual embeddings and the target linguistic space. This random choice has negligible effect on final quality (see Appendix). During this stage (as illustrated in Fig. 2, where the flow of this stage is indicated by solid and dashed lines), the sign language backbones remain frozen, while the adapters, routing and fusion modules, and the LLM encoder are trainable; the LLM decoder is kept frozen to preserve its linguistic generation priors. Let ψ_v denote the LLM encoder that processes fused visual embeddings, ψ_d the LLM decoder, and ψ_t the frozen text encoder used for generating textual embeddings.

Given the fused visual embedding $\mathcal{Z}_{\text{fused}}$, the LLM encoder produces a global visual-language representation $\mathbf{v} = \text{MeanPool}(\psi_v(\mathcal{Z}_{\text{fused}})) \in \mathbb{R}^H$. Similarly, for each ground-truth sentence $\mathbf{y} \in \mathbf{y}^*$, its textual embedding is computed using the frozen language encoder as $\mathbf{t} = \text{MeanPool}(\psi_t(\mathbf{y})) \in \mathbb{R}^H$. We employ a symmetric InfoNCE loss to align the visual and textual embeddings:

$$\mathcal{L}_{\text{con}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(\mathbf{v}_i^\top \mathbf{t}_i/\tau_c)}{\sum_{j=1}^{B}\exp(\mathbf{v}_i^\top \mathbf{t}_j/\tau_c)} + \log\frac{\exp(\mathbf{t}_i^\top \mathbf{v}_i/\tau_c)}{\sum_{j=1}^{B}\exp(\mathbf{t}_i^\top \mathbf{v}_j/\tau_c)}\right] \tag{9}$$

where τ_c is a learned temperature parameter controlling distribution sharpness and B is the batch size. This alignment stage grounds the visual encoder in the textual embedding space, enabling robust adaptation to new languages before final fine-tuning.

3.8 Stage III: Fine-tuning on Downstream Datasets

During fine-tuning (as illustrated in Fig. 2, where the flow of this stage is indicated by solid and dotted lines), the pretrained sign language backbones remain

frozen to preserve the robust visual-linguistic representations learned in earlier stages, while the adapters, attention-based fusion, gating modules, and the LLM encoder-decoder are updated. Given the skeletal keypoints extracted from each downstream video, the frozen backbones generate part-specific features that are aggregated through attention-guided fusion into the unified embedding $\mathcal{Z}_{\text{fused}}$. This embedding is passed to the LLM encoder, which encodes visual-semantic information, and subsequently to the decoder ψ_d to autoregressively generate the target translation. The model is optimised using:

$$\mathcal{L}_{\text{fine-tune}} = -\sum_{t=1}^{T''} \log P(y_t \mid y_{<t}, \mathcal{Z}_{\text{fused}}), \tag{10}$$

where $\mathbf{y} = (y_1, \ldots, y_{T''}) \in \mathbf{y}^*$ denotes the target text sequence. This stage enables efficient adaptation to new sign languages without compromising the strong priors established during large-scale pretraining.

4 Experiments

Large-Scale Datasets. We pretrain on three large-scale, sentence-aligned video-text datasets: CSL-News [31] (1,985 hours) for Chinese Sign Language, YT-ASL [48] (1,394 hours) for American Sign Language, and a BSL corpus constructed by combining the BSL Corpus [38, 39] (125 hours), the BSL subset of YT-SL25 [44] (74 hours), and BOBSL [2] (1,467 hours). Each dataset trains a dedicated backbone jointly with an LLM encoder-decoder, yielding three pretrained experts (N=3) within our routing and fusion architecture.

Downstream Datasets. We evaluate our model on four gloss-free SLT benchmarks spanning diverse sign languages and resource levels: How2Sign [14] (79 hours, ASL) contains aligned video-text pairs of American Sign Language; CSL-Daily [59] (23 hours, CSL) features conversational Chinese Sign Language across daily topics; Phoenix14T [6] (11 hours, DGS) offers German Sign Language weather broadcasts with precise alignments; and MeineDGS [21] (50 hours, DGS) provides a linguistically rich corpus with high signer and vocabulary diversity.

Metrics. We evaluate translation performance using BLEU-1 to BLEU-4 and ROUGE-L scores, which measure n-gram precision and recall with an emphasis on longest common subsequences [5, 33, 36]. Additionally, BLEURT is employed to assess semantic adequacy and fluency, capturing meaning similarity beyond lexical overlap [40]. We report per-instance (P-I) and per-class (P-C) Top-1 accuracy for evaluating Sign Language Recognition (SLR).

Implementation Details. Our training pipeline consists of three stages: large-scale pretraining, contrastive alignment, and downstream fine-tuning. We use mT5-base as the multilingual LLM encoder-decoder across all three stages. For Stage I, each large-scale dataset is used to train a dedicated sign language backbone together with its LLM encoder-decoder in an end-to-end manner for 20 epochs. Stage II is also trained for 20 epochs to adapt representations for unseen languages, and Stage III runs for 40 epochs on each downstream dataset. See Appendix for more details.

Table 1: SLT results on Phoenix14T and CSL-Daily. BLEU-1, BLEU-4, and ROUGE-L are abbreviated as B-1, B-4, and R.

Method	Modality		Phoenix14T						CSL-Daily					
			Dev			Test			Dev			Test		
	Pose	RGB	B-1↑	B-4↑	R↑	B-1↑	B-4↑	R↑	B-1↑	B-4↑	R↑	B-1↑	B-4↑	R↑
Gloss-based														
SLTUNET [55]		✓	–	27.87	52.23	52.92	28.47	52.11	–	23.99	53.58	54.98	25.01	54.08
MMTLB [8]		✓	53.95	27.61	53.10	53.97	28.39	52.65	53.81	24.42	53.38	53.31	23.92	53.25
IP-SLT [52]		✓	54.10	28.22	54.43	54.25	27.97	53.72	45.26	16.74	44.33	44.85	16.72	44.09
TS-SLT [9]	✓	✓	54.32	28.66	54.08	54.90	28.95	53.48	55.21	25.76	55.10	55.44	25.79	55.72
Gloss-free														
GFSLT-VLP [57]		✓	44.08	22.12	43.72	43.71	21.44	42.49	39.20	11.07	36.70	39.37	11.00	36.44
FLa-LLM [11]		✓	–	–	–	46.29	23.09	45.27	–	–	–	37.13	14.20	37.25
Sign2GPT [50]		✓	–	–	–	49.54	22.52	48.90	–	–	–	41.75	15.40	42.36
SignLLM [17]		✓	46.88	25.25	47.23	45.21	23.40	44.49	42.45	12.23	39.18	39.55	15.75	39.91
MMSLT [28]		✓	48.73	25.47	48.58	48.92	25.73	47.97	50.05	20.51	48.53	49.87	21.11	48.92
BeyondGloss [4]		✓	–	–	–	52.38	25.49	52.89	–	–	–	53.12	21.53	53.46
PGG-SLT [20]		✓	53.84	27.53	52.85	54.02	27.32	52.56	–	–	–	–	–	–
Uni-Sign [31]	✓	✓	–	–	–	–	–	–	55.30	26.25	56.03	55.08	26.36	56.51
Geo-Sign [16]	✓		–	–	–	–	–	–	55.57	27.05	57.27	55.89	27.42	57.95
SIGNET	✓		**54.15**	**27.78**	**53.01**	**54.10**	**27.82**	**53.05**	**56.90**	**28.27**	**58.48**	**56.77**	**28.51**	**58.66**

4.1 Comparison with Prior Methods

Results on Phoenix14T. As seen in Tab. 1, **SIGNET** achieves state-of-the-art results among gloss-free methods using only skeletal keypoints, without requiring RGB frames or gloss annotations. Compared to PGG-SLT [20], which leverages an LLM to predict and order glosses through prompt-based reasoning, **SIGNET** achieves a +0.50 BLEU-4 gain with a substantially lighter input representation, highlighting the effectiveness of multi-expert fusion for cross-lingual transfer.

Results on CSL-Daily. Based on Tab. 1, on this dataset, our method achieves clear gains over prior approaches, outperforming RGB-based models by +6.9 BLEU-4 and showing consistent improvements over pose-based methods, with a +1.09 BLEU-4 increase compared to Geo-Sign [16]. While Geo-Sign employs computationally intensive geometric operations that introduce higher latency during training [16], our framework attains superior efficiency through a lightweight adaptation mechanism that scales across multiple pretrained backbones.

Results on How2Sign. Tab. 2 highlights that our model achieves state-of-the-art performance on How2Sign, outperforming the pose-based Uni-Sign [31] and Geo-Sign [16], and comparable results to the RGB-based SSVP-SLT-LSP [37]. However, this method requires an extremely demanding training setup, utilising 64 A100 GPUs for two weeks and leveraging both YT-ASL and How2Sign during pretraining as well as during fine-tuning, making it difficult to extend to additional datasets [37]. In contrast, our approach achieves these results through an efficient and scalable design that only requires fine-tuning on How2Sign.

Results on MeineDGS. Tab. 3 shows that **SIGNET** generalises to MeineDGS, a dataset whose language (DGS) was not seen during pretraining, showcasing the transferability of motion-level priors to new sign languages. Unlike Sincan

Table 2: SLT results on How2Sign. † denotes pretraining on YT-ASL and How2Sign. BLEU-1, BLEU-4, ROUGE-L and BLEURT are abbreviated as B-1, B-4, R and B-RT.

Method	Modality		Test			
	Pose	RGB	B-1↑	B-4↑	R↑	B-RT↑
Gloss-free						
SSVP-SLT-LSP† [37]		✓	43.2	15.5	38.4	49.6
YouTube-ASL [48]	✓		37.8	12.4	–	46.6
SLT-IV [45]		✓	34.0	8.0	–	–
C^2RL [10]		✓	29.1	9.4	27.0	–
FLa-LLM [11]		✓	29.8	9.7	27.8	–
PGG-SLT [20]		✓	40.8	13.7	32.9	–
Jang et al. [25]		✓	–	12.7	32.5	45.3
SignMusketeers [18]		✓	41.5	14.3	–	–
SSVP-SLT [37]		✓	41.9	14.7	37.8	49.3
Uni-Sign [31]	✓	✓	40.2	14.9	36.0	49.4
Geo-Sign [16]	✓		40.8	15.1	35.4	–
SIGNET	✓		**41.1**	**15.4**	**36.4**	**49.5**

Table 3: SLT results on MeineDGS.

Method	Modality		Test			
	Pose	RGB	B-1↑	B-4↑	R↑	B-RT↑
Gloss-free						
Spotter+GPT [42]		✓	14.82	0.64	–	21.62
Spotter+Transformer [42]		✓	19.50	1.08	–	19.01
Sub-GT+GPT [42]		✓	16.65	1.55	–	29.72
SIGNET	✓		**24.20**	**2.75**	**13.6**	**33.58**

Table 4: SLR results on WLASL. P-I and P-C denote per-instance and per-class Top-1 accuracy, respectively.

Method	Modality		WLASL100		WLASL2000	
	Pose	RGB	P-I↑	P-C↑	P-I↑	P-C↑
NLA-SLR [63]	✓	✓	91.47	92.17	61.05	58.05
SignRep [49]		✓	–	–	61.05	58.89
Uni-Sign [31]	✓	✓	92.25	92.67	63.52	61.32
Geo-Sign [16]	✓		–	–	63.64	61.89
SIGNET	✓		**93.26**	**93.47**	**64.87**	**62.07**

et al. [42], which partially leverage ground-truth glosses and GPT, our method operates in a fully gloss-free manner while achieving better performance.

Results on WLASL. We evaluate **SIGNET** for isolated sign recognition on WLASL100 and WLASL2000 [29] to assess the transferability of learned visual-linguistic features. Without model modifications, we skip Stage II and replace target sentences with individual glosses in Stage III training. As shown in Tab. 4, **SIGNET** outperforms existing pose-based and RGB-based methods, suggesting that representations learned by **SIGNET** effectively capture transferable sign semantics beyond translation tasks.

Computational Efficiency. Tab. 5 compares parameter counts and forward latency across methods. Despite using multiple pretrained backbones, **SIGNET** requires only 1.5M trainable parameters in the video encoder, as backbone features can be pre-computed and cached. Stage I pretraining is a one-time cost (576 GPU-hours, 4×RTX 3090); downstream adaptation requires only ∼24 GPU-hours on a single RTX 3090. In comparison, SSVP-SLT-LSP demands 21,504 A100-hours for a single language, yet **SIGNET** matches its How2Sign performance (15.4 vs. 15.5 B-4) using 64× fewer GPUs without requiring How2Sign during pretraining. **SIGNET** achieves a per-batch training latency of 320 ms, compared to Uni-Sign (416 ms) and Geo-Sign (2550 ms), both of which train and evaluate on a single language and do not address cross-lingual transfer.

Qualitative Results. See Appendix for qualitative examples.

4.2 Ablation Study

Cross-lingual Transfer Benefits. Tab. 6 compares performance when finetuning with trainable versus frozen backbones pretrained on large-scale datasets. For this evaluation, we pretrain a sign language backbone and an LLM on a large-scale dataset (similar to Sec. 3.3) and without adding any extra modules,

Table 5: Computational efficiency comparison. Parameters in millions (M); forward latency in milliseconds (ms) per training batch (batch size 8). †Backbone features can be pre-computed and cached after Stage I. The reported latency is *non-cached* (all three backbones run per batch).

Method	Video Encoder		Language Model		Full Model		
	#Total	#Train	#Total	#Train	#Total	#Train	Fwd Latency
RGB-based Methods							
Sign2GPT [50]	34.9M	12.8M	1736.7M	3.8M	1771.6M	16.7M	–
C^2RL [10]	11.7M	11.7M	680M	680M	691.7M	691.7M	–
Skeleton-based Methods							
Uni-Sign [31]	9.7M	9.7M	582.4M	582.4M	592.1M	592.1M	416 ms
Geo-Sign [16]	6.7M	6.7M	582.4M	582.4M	589.1M	589.1M	2550 ms
SIGNET	17.4M†	**1.5M**	582.4M	582.4M	599.8M	583.9M	**320 ms**

we fine-tune it directly on the downstream datasets after loading the pretrained weights. This table confirms the assumptions that our paper is based upon and that adapting frozen backbones yield substantially higher performance across cross-lingual settings, while trainable backbones suffer from overfitting and catastrophic forgetting. Notably, when the pretraining and downstream datasets share the same language, both settings perform comparably. However, in cross-lingual transfer (e.g., CSL-News → Phoenix14T or YT-ASL → CSL-Daily), the frozen backbone consistently achieves stronger results, with some scores comparable to prior methods. These findings suggest that pretrained backbones retain transferable articulatory representations across sign languages, supporting our design of keeping them frozen and adapting through lightweight modules.

Feature Aggregation Input. In Tab. 7, we observe that using only the hand features as input priors to the feature aggregation module yields the best result across all datasets, while incorporating body and facial priors reduces translation quality. We attribute this to the attention-based aggregation being optimised for compact, semantically dense inputs; the higher-dimensional body and face features likely introduce noise into the routing signal. This is consistent with the role of hands as the primary carriers of lexical and semantic content in sign languages, making them the most effective cue to distinguish expert contributions.

Components Contributions. Tab. 8 presents the effect of individual components and backbone configurations on translation performance. Specifically, removing Gating replaces learned routing with uniform averaging, removing Adapter discards the low-rank residual updates, and removing Contrastive skips Stage II alignment. We systematically evaluate the contribution of the *Adapter*, *Gating*, and *Contrastive* modules under different numbers of pretrained backbones ($N \in \{1, 2, 3\}$) and routing top-k selections. On Phoenix14T, the best performance is achieved when using three backbones with $k = 2$ and all components enabled, demonstrating that multi-expert fusion and contrastive loss jointly enhance cross-modal correspondence. Interestingly, for CSL-Daily, removing the contrastive loss slightly improves results, suggesting that excessive cross-lingual

Table 6: Adaptation impact on addressing overfitting. ↑/↓: B-4 change from freezing.

Pretraining Dataset	Phoenix14T	CSL-Daily	How2Sign
	B-4↑	B-4↑	B-4↑
Trainable Backbone			
CSL-News	11.47	**25.56**	6.8
YT-ASL	12.04	13.8	**14.4**
Frozen Backbone			
CSL-News	**23.30** ↑11.8	24.65 ↓0.9	**10.3** ↑3.5
YT-ASL	**24.05** ↑12.0	**20.7** ↑6.9	13.1 ↓1.3

Table 7: Impact of different input configurations for feature aggregation on performance.

Phoenix14T			CSL-Daily		
B-1↑	B-4↑	R↑	B-1↑	B-4↑	R↑
Hands					
54.15	**27.78**	**53.01**	**56.90**	**28.27**	**58.48**
Hands + Body					
53.83	26.23	50.81	54.67	26.78	56.45
Hands + Body + Face					
51.11	25.12	48.08	52.56	24.94	53.64

Table 8: Ablation on top-k gating, components, and backbone count. **(i)** No Gating replaces learned routing weights with uniform averaging over experts; **(ii)** No Adapter removes the low-rank residual updates while keeping frozen expert features; **(iii)** No Contrastive skips Stage II alignment and proceeds directly to downstream fine-tuning.

k	Components			Phoenix14T		CSL-Daily	
	Adapter	Gating	Contrastive	B-4↑	R↑	B-4↑	R↑
1 Backbone							
–	✗	✗	✗	24.05	47.52	25.56	53.82
–	✓	✗	✗	23.85	46.60	26.23	55.45
–	✗	✗	✓	24.65	46.62	26.20	54.44
–	✓	✗	✓	25.12	47.78	25.58	53.95
2 Backbones							
1	✗	✓	✗	25.36	47.88	24.45	53.92
1	✓	✓	✗	26.35	49.48	27.30	56.12
1	✓	✓	✓	27.15	51.08	26.65	55.12
2	✓	✓	✓	27.10	51.04	26.10	54.23
3 Backbones							
1	✓	✓	✓	27.20	50.98	27.25	55.62
2	✓	✓	✓	**27.78**	**53.01**	26.50	55.84
2	✓	✓	✗	27.63	52.87	**28.27**	**58.48**
3	✓	✓	✓	26.30	49.54	26.90	55.52

regularisation may interfere with datasets that already share language overlap with pretraining corpora. Overall, the results highlight the complementary role of each component, *Adapters* facilitate efficient transfer, *Gating* improves dynamic expert selection, and *Contrastive alignment* strengthens visual-linguistic grounding when domain or language gaps are present. With three pretrained backbones, several combinations are possible when selecting one or two experts. The reported results correspond to the best configuration.

Effect of k on SLT. According to Tab. 8, performance does not increase monotonically with k. The best results arise with three backbones and k=2. Using a single expert (k=1) limits cues, while using all experts (k=3) introduces redundancy and over-smoothing. k=2 strikes a sparsity–diversity trade-off, combining two complementary experts while keeping the mixture selective. Empirically, this setting yields higher gate confidence and lower pairwise expert redundancy.

Expert Contribution Analysis. To understand how motion-level knowledge transfers across languages, we remove one expert at a time from the full three-backbone configuration and report the resulting BLEU-4 drop in Tab. 9. When the downstream dataset shares a language with a pretrained backbone, removing that backbone causes the largest performance drop. On CSL-Daily, removing the CSL expert leads to a substantial 6.56 BLEU-4 decrease, far exceeding the impact of removing ASL (1.48) or BSL (0.43). Similarly, on How2Sign, removing the ASL expert causes a 5.6-point drop, while the other removals have minimal effect (0.5 and 1.3). These results confirm that the framework effectively identifies and relies on the language-matched expert when one is available. For the DGS

Table 9: Leave-one-expert-out analysis (k=2). Δ denotes the BLEU-4 drop when each pretrained backbone is removed from the full three-expert model; the largest drop per dataset is **bolded**. Per dataset, green / red shade higher BLEU-4 / larger drop.

	Phoenix14T		CSL-Daily		How2Sign		MeineDGS	
	B-4↑	Δ	B-4↑	Δ	B-4↑	Δ	B-4↑	Δ
SIGNET	**27.78**	–	**28.27**	–	**15.4**	–	**2.75**	–
w/o **ASL expert**	23.04	**4.74**	26.79	1.48	9.8	**5.6**	1.05	**1.7**
w/o **BSL expert**	24.02	3.76	27.84	0.43	14.9	0.5	1.65	1.1
w/o **CSL expert**	24.68	3.10	21.71	**6.56**	14.1	1.3	2.05	0.7

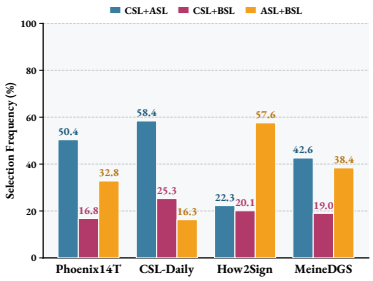

Fig. 4: Per-dataset expert pair selection (k=2).

datasets, where no matching backbone was seen during pretraining, the pattern is more nuanced. On Phoenix14T, the ASL expert is the most critical (4.74 drop), followed by BSL (3.76) and CSL (3.10), indicating that all three experts contribute meaningfully and that transfer to DGS draws on complementary cues rather than a single dominant source. On MeineDGS, a similar trend holds with ASL again being the most impactful (1.7 drop), while BSL and CSL contribute comparably (1.1 and 0.7). These drop patterns are consistent with the expert pair selection frequencies shown in Fig. 4. On CSL-Daily and How2Sign, the two pairs containing the language-matched backbone together account for 83.7% (CSL+ASL and CSL+BSL) and 79.9% (CSL+ASL and ASL+BSL) of selections, respectively, whereas on the DGS datasets the selection is more distributed, with no single pair exceeding 51%. The asymmetric drop patterns indicate language-specific rather than generic temporal representations, with expert importance shaped by linguistic relatedness and dataset properties. The language-matched backbone dominates when available, while for languages absent from Stage I pretraining, the framework distributes reliance across experts, explaining the strong cross-lingual transfer in our main results.

Impact of Pretraining Data Scale. To examine how pretraining data volume affects downstream translation, we randomly subsample varying proportions (25%, 50%, 75%, 100%) from each of the three pretraining datasets (CSL, ASL, and BSL) for Stage I. As shown in Fig. 5, translation performance improves consistently as the pretraining data increases, and notably the trend does not plateau at full scale, indicating that the learned representations are not yet saturated and that the bottleneck lies in pretraining coverage rather than in the adaptation modules. This confirms that **SIGNET** benefits from larger-scale pretraining and suggests further gains as additional sign language data become available.

SIGNET vs. Naive Combination. As shown in Tab. 10, simply combining multiple pretrained backbones via concatenation or averaging assumes uniform relevance and increases overfitting risk. Averaging treats every expert as equally informative regardless of the input, diluting the language-matched backbone, while concatenation inflates the input dimensionality and provides more oppor-

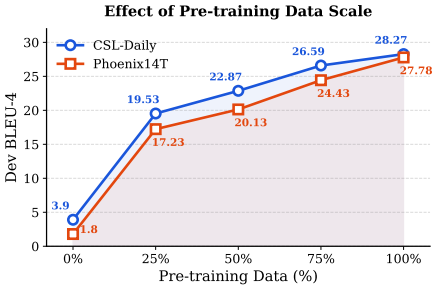

Fig. 5: Pretraining Data Scale Impact.

Table 10: Fusion method comparison. BLEU-4 on Phoenix14T and CSL-Daily for combining the three pretrained experts via uniform averaging, feature concatenation, or our input-dependent adaptive gating.

	Phoenix14T	CSL-Daily
Fusion Method	B-4↑	B-4↑
Average all experts ($\frac{1}{N}$)	24.12	25.08
Concatenate experts	24.30	25.75
SIGNET(adaptive gating)	**27.78**	**28.27**

tunities to overfit the limited downstream data. Both underperform. Averaging yields 24.12 and 25.08 BLEU-4 on Phoenix14T and CSL-Daily, respectively, while concatenation achieves 24.30 and 25.75. In contrast, our adaptive gating conditions the fusion on each input, sparsely activating only the most relevant experts, which preserves expert specialisation while limiting parameter growth. This substantially outperforms both static baselines (+3.5 BLEU-4 on average), showing that learned sparse routing effectively balances sparsity and diversity.

5 Conclusion

We presented **SIGNET**, a framework for motion-level knowledge transfer across sign languages that treats independently pretrained backbones as frozen visual experts with distinct motion-level specialisations. Through hand-prior-driven routing and lightweight adaptation, **SIGNET** integrates multiple experts using only 1.5M trainable video encoder parameters while matching methods requiring orders of magnitude more compute. Our analysis reveals asymmetric expert reliance patterns consistent with language-specific rather than generic representations, with the framework automatically discovering complementary expertise for languages absent from pretraining. Comprehensive experiments demonstrate state-of-the-art translation performance among gloss-free methods across multiple datasets, while also surpassing prior methods on sign language recognition.

Acknowledgements

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

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing systems (NeurIPS) **35**, 23716–23736 (2022)
2. 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. arXiv preprint arXiv:2111.03635 (2021), `https://www.robots.ox.ac.uk/~vgg/data/bobsl`
3. Asasi, S., Lakhal, M.I., Bowden, R.: Hierarchical feature alignment for gloss-free sign language translation. In: International Conference on Intelligent Virtual Agents (IVA Adjunct). Association for Computing Machinery (ACM) (2025)
4. Asasi, S., Lakhal, M.I., Sincan, O.M., Bowden, R.: Beyond gloss: A hand-centric framework for gloss-free sign language translation. In: British Machine Vision Conference (BMVC). British Machine Vision Association (2025)
5. Brown, P.F., Della Pietra, V.J., Desouza, P.V., Lai, J.C., Mercer, R.L.: Class-based n-gram models of natural language. Computational linguistics **18**(4), 467–480 (1992)
6. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7784–7793 (2018)
7. Chen, Y., Wei, F., Sun, X., Wu, Z., Lin, S.: 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). pp. 5120–5130 (2022)
8. Chen, Y., Wei, F., Sun, X., Wu, Z., Lin, S.: 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). pp. 5120–5130 (2022)
9. Chen, Y., Zuo, R., Wei, F., Wu, Y., Liu, S., Mak, B.: Two-stream network for sign language recognition and translation. Advances in Neural Information Processing systems (NeurIPS) **35**, 17043–17056 (2022)
10. Chen, Z., Zhou, B., Huang, Y., Wan, J., Hu, Y., Shi, H., Liang, Y., Lei, Z., Zhang, D.: 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 (2025)
11. 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: Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING). pp. 7071–7081 (2024)
12. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in Neural Information Processing systems (NeurIPS) **36**, 49250–49267 (2023)
13. Du, N., Huang, Y., Dai, A.M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A.W., Firat, O., et al.: Glam: Efficient scaling of language models with mixture-of-experts. In: International Conference on Machine Learning (ICML). pp. 5547–5569. PMLR (2022)

14. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2735–2744 (2021)
15. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research (JMLR) **23**(120), 1–39 (2022)
16. Fish, E., Bowden, R.: Geo-sign: Hyperbolic contrastive regularisation for geometrically aware sign language translation. Advances in Neural Information Processing systems (NeurIPS) **38**, 99293–99330 (2026)
17. Gong, J., Foo, L.G., He, Y., Rahmani, H., Liu, J.: Llms are good sign language translators. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18362–18372 (2024)
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. pp. 22506–22521 (2025)
19. Guo, J., Cai, Y., Bi, K., Fan, Y., Chen, W., Zhang, R., Cheng, X.: Came: Competitively learning a mixture-of-experts model for first-stage retrieval. ACM Transactions on Information Systems **43**(2), 1–25 (2025)
20. Guo, J., Li, P., Cohn, T.: Bridging sign and spoken languages: Pseudo gloss generation for sign language translation. Advances in Neural Information Processing systems (NeurIPS) **38**, 77471–77499 (2026)
21. Hanke, T., König, S., Konrad, R., Langer, G., Barbeito Rey-Geißler, P., Blanck, D., Goldschmidt, S., Hofmann, I., Hong, S.E., Jeziorski, O., Kleyboldt, T., König, L., Matthes, S., Nishio, R., Rathmann, C., Salden, U., Wagner, S., Worseck, S.: MEINE DGS. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release (2020). `https://doi.org/10.25592/dgs.meinedgs-3.0`, `https://doi.org/10.25592/dgs.meinedgs-3.0`
22. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning (ICML). pp. 2790–2799. PMLR (2019)
23. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. International Conference on Learning Representations (ICLR) (2022)
24. Jang, E., Gu, S., Poole, B.: Categorical reparametrization with gumble-softmax. In: International Conference on Learning Representations (ICLR) (2017)
25. Jang, Y., Raajesh, H., Momeni, L., Varol, G., Zisserman, A.: Lost in translation, found in context: Sign language translation with contextual cues. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8742–8752 (2025)
26. Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., Chen, K.: Rtmpose: Real-time multi-person pose estimation based on mmpose. arXiv preprint arXiv:2303.07399 (2023)
27. Karimi Mahabadi, R., Henderson, J., Ruder, S.: Compacter: Efficient low-rank hypercomplex adapter layers. Advances in Neural Information Processing systems (NeurIPS) **34**, 1022–1035 (2021)
28. Kim, J., Jeon, H., Bae, J., Kim, H.Y.: 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 (2025)

29. Li, D., Rodriguez, C., Yu, X., Li, H.: Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1459–1469 (2020)
30. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning (ICML). pp. 19730–19742. PMLR (2023)
31. Li, Z., Zhou, W., Zhao, W., Wu, K., Hu, H., Li, H.: Uni-sign: Toward unified sign language understanding at scale. In: International Conference on Learning Representations (ICLR) (2025)
32. 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)
33. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004)
34. Ma, N., Zhang, H., Li, X., Zhou, S., Zhang, Z., Wen, J., Li, H., Gu, J., Bu, J.: Learning spatial-preserved skeleton representations for few-shot action recognition. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 174–191. Springer (2022)
35. Maddison, C., Mnih, A., Teh, Y.: The concrete distribution: A continuous relaxation of discrete random variables. In: International Conference on Learning Representations (ICLR) (2017)
36. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). p. 311–318. ACL '02, Association for Computational Linguistics, USA (2002)
37. Rust, P., Shi, B., Wang, S., Camgöz, N.C., Maillard, J.: Towards privacy-aware sign language translation at scale. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). pp. 8624–8641 (2024)
38. Schembri, A., Fenlon, J., Rentelis, R., Reynolds, S., Cormier, K.: Building the british sign language corpus. University of Hawaii Press (2013)
39. Schembri, A., Jordan, F., Rentelis, R., Cormier, K.: British sign language corpus project: A corpus of digital video data and annotations of british sign language 2008-2017 (third edition). (2017), `https://www.bslcorpusproject.org`
40. Sellam, T., Das, D., Parikh, A.: Bleurt: Learning robust metrics for text generation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). pp. 7881–7892 (2020)
41. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: International Conference on Learning Representations (ICLR) (2017)
42. Sincan, O.M., Bowden, R.: Spotter+ gpt: Turning sign spottings into sentences with llms. In: International Conference on Intelligent Virtual Agents (IVA Adjunct). No. In Press, Association for Computing Machinery (ACM) (2025)
43. Sincan, O.M., Low, J.H., Asasi, S., Bowden, R.: Gloss-free sign language translation: An unbiased evaluation of progress in the field. Computer Vision and Image Understanding p. 104498 (2025)
44. Tanzer, G., Zhang, B.: Youtube-sl-25: A large-scale, open-domain multilingual sign language parallel corpus. In: International Conference on Learning Representations (ICLR). vol. 2025, pp. 81921–81934 (2025)

45. Tarrés, L., Gállego, G.I., Duarte, A., Torres, J., Giró-i Nieto, X.: Sign language translation from instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5625–5635 (2023)
46. Thatipelli, A., Narayan, S., Khan, S., Anwer, R.M., Khan, F.S., Ghanem, B.: Spatio-temporal relation modeling for few-shot action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19958–19967 (2022)
47. Thomas, M., Fish, E., Bowden, R.: Vallr: Visual asr language model for lip reading. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2846–2856 (2025)
48. Uthus, D., Tanzer, G., Georg, M.: Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. Advances in Neural Information Processing systems (NeurIPS) **36**, 29029–29047 (2023)
49. Wong, R., Camgoz, N.C., Bowden, R.: Signrep: Enhancing self-supervised sign representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22804–22814 (2025)
50. Wong, R.C., Camgöz, N.C., Bowden, R.: Sign2gpt: leveraging large language models for gloss-free sign language translation. In: International Conference on Learning Representations (ICLR) (2024)
51. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
52. Yao, H., Zhou, W., Feng, H., Hu, H., Zhou, H., Li, H.: Sign language translation with iterative prototype. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15592–15601 (2023)
53. Ye, J., Wang, X., Jiao, W., Liang, J., Xiong, H.: Improving gloss-free sign language translation by reducing representation density. Advances in Neural Information Processing systems (NeurIPS) **37**, 107379–107402 (2024)
54. Yin, A., Zhao, Z., Liu, J., Jin, W., Zhang, M., Zeng, X., He, X.: Simulslt: End-to-end simultaneous sign language translation. In: Proceedings of the ACM International Conference on Multimedia (MM). pp. 4118–4127 (2021)
55. Zhang, B., Müller, M., Sennrich, R.: Sltunet: A simple unified model for sign language translation. In: International Conference on Learning Representations (ICLR) (2023)
56. Zhang, H., Zhang, L., Qi, X., Li, H., Torr, P.H., Koniusz, P.: Few-shot action recognition with permutation-invariant attention. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 525–542. Springer (2020)
57. 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). pp. 20871–20881 (2023)
58. Zhou, H., Wang, Z., Huang, S., Huang, X., Han, X., Feng, J., Deng, C., Luo, W., Chen, J.: Moe-lpr: Multilingual extension of large language models through mixture-of-experts with language priors routing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 26092–26100 (2025)
59. Zhou, H., Zhou, W., Qi, W., Pu, J., Li, H.: 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). pp. 1316–1325. IEEE (2021)

60. Zhou, H., Zhou, W., Qi, W., Pu, J., Li, H.: 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). pp. 1316–1325 (2021)
61. Zhou, H., Zhou, W., Zhou, Y., Li, H.: Spatial-temporal multi-cue network for sign language recognition and translation. IEEE Transactions on Multimedia (TMM) **24**, 768–779 (2021)
62. Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V., Dai, A.M., Le, Q.V., Laudon, J., et al.: Mixture-of-experts with expert choice routing. Advances in Neural Information Processing systems (NeurIPS) **35**, 7103–7114 (2022)
63. Zuo, R., Wei, F., Mak, B.: Natural language-assisted sign language recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14890–14900 (2023)