SAGE: Segment-Aware Gloss-Free Encoding for Token-Efficient Sign Language Translation

Low Jian He Ozge Mercanoglu Sincan Richard Bowden
CVSSP, University of Surrey, United Kingdom
{jianhe.low, o.mercanoglusincan, r.bowden}@surrey.ac.uk

Abstract

Gloss-free Sign Language Translation (SLT) has advanced rapidly, achieving strong performances without relying on gloss annotations. However, these gains have often come with increased model complexity and high computational demands, raising concerns about scalability, especially as large-scale sign language datasets become more common.

We propose a segment-aware visual tokenization framework that leverages sign segmentation to convert continuous video into discrete, sign-informed visual tokens. This reduces input sequence length by up to 50% compared to prior methods, resulting in up to 2.67× lower memory usage and better scalability on larger datasets. To bridge the visual and linguistic modalities, we introduce a token-to-token contrastive alignment objective, along with a dual-level supervision that aligns both language embeddings and intermediate hidden states. This improves fine-grained cross-modal alignment without relying on gloss-level supervision. Our approach notably exceeds the performance of state-of-the-art methods on the PHOENIX14T benchmark, while significantly reducing sequence length. Further experiments also demonstrate our improved performance over prior work under comparable sequence-lengths, validating the potential of our tokenization and alignment strategies.

1. Introduction

Sign languages are the primary means of communication within Deaf communities and are natural human languages with rich linguistic structure and deep historical roots [31]. They go beyond simple visual representations of speech, possessing their own phonological, morphological, and syntactic systems [23, 32], distinct from spoken languages.

While recent advances in deep learning have transformed natural language understanding, most progress has been driven by text and speech modalities, supported by decades of research and large-scale resources. Sign language, in contrast, is inherently visual and multimodal, unfolding across space and time through coordinated hand gestures, facial expressions, and body movements [19]. These multimodal signals interact in a continuous, fine-grained manner, making the modeling task especially difficult as no single modality or frame captures the full linguistic content.

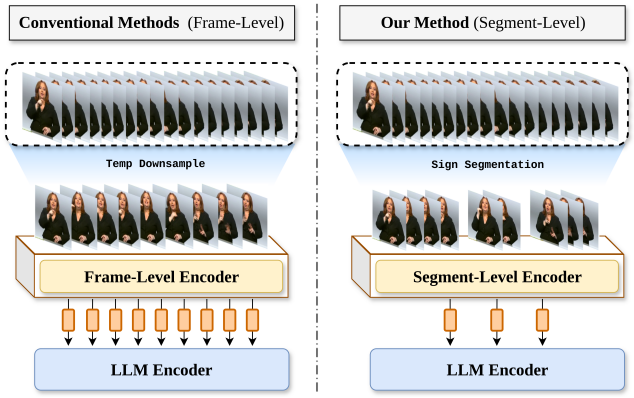

Figure 1. **Method Comparison.** Conventional approaches rely on temporal downsampling to reduce the number of video frames; however, this often still results in lengthy and redundant sequences being passed to the language encoder. In contrast, our method introduces a visual tokenization strategy that groups frames into discrete visual tokens, significantly reducing input sequence length.

Sign Language Translation (SLT), the task of translating continuous sign videos into spoken language sentences, is therefore notably challenging as it requires models to bridge complex spatiotemporal input and structured language output. A common strategy in prior work is to use gloss annotations as intermediate supervision [4, 6, 41, 42, 46]. Glosses are written representations of signs and thus provide explicit labels to guide models in recognizing signs before translation. While effective, gloss annotations are extremely costly [5], as they require frame-level labeling by expert linguists, creating a major bottleneck for scaling SLT systems.

To overcome this, recent efforts have shifted towards gloss-free SLT [3, 12, 37, 45], which bypasses intermediate supervision and trains directly on paired sign videos

and spoken language sentences. These approaches often leverage contrastive learning, employing CLIP-style objectives [34, 45] or frame-level contrastive losses [39], to align visual and textual embeddings. However, without explicit gloss labels, the supervision signal becomes weaker. To compensate, some works rely on large-scale pretraining [11, 22], while others attempt to impose gloss-like alignment [37]. In all cases, the absence of glosses leads to a heavier dependence on larger and more expressive architectures to learn the intricate cross-modal correspondences.

Additionally, the inherent fine-grained nature of sign language further exacerbates the modeling challenge. Sign language encodes meaning through subtle, continuous changes; thus, even slight variations can alter the semantic content, necessitating dense, frame-level processing to preserve linguistic fidelity. As recent SLT approaches have predominantly adopted Transformer-based visual encoders, full-resolution temporal encoding becomes computationally infeasible as the self-attention mechanism in Transformers scales quadratically with sequence length [26]. To manage this, most approaches have introduced temporal reduction strategies, such as max pooling [45], uniform downsampling [8, 37], and sliding window mechanisms [12]. However, these methods overlook the linguistic structure of sign language, treating all frames as equally informative. This is problematic, as sign language exhibits variable sign duration and is sensitive to individual signing speeds [2]. Thus, agnostic reduction risks discarding semantically critical frames, ultimately constraining a model's ability to capture the expressive richness of sign language.

Motivated by these challenges, we propose a *segment-aware gloss-free encoding* framework for efficient SLT. As discussed, existing methods predominantly rely on frame-level operations; however, sign language segmentation, despite its inherent ability to group frames into coherent sign units, remains underutilized. Our implementation therefore explores leveraging a segmentation model [25] as a visual tokenizer to localize meaningful sign units within continuous signing. Instead of processing dense sequences, we aggregate frames into semantically coherent segments and encode each segment as a single visual token. This significantly reduces sequence length, alleviates the computational burden associated with Transformer-based encoders, and more faithfully captures the linguistic structure of sign language. As shown in Fig. 1, our method yields significantly fewer tokens than frame-based downsampling. To align visual tokens with language, we further introduce a *token-level contrastive alignment objective* that encourages fine-grained correspondence between visual and text tokens. In summary, our main contributions are as follows:

- We propose the first gloss-free SLT framework that incorporates sign language segmentation to produce semantically aligned, segment-based visual tokens.
- Our approach achieves substantial temporal reduction, significantly lowering token count and Transformer computation while maintaining translation quality.
- We introduce a token-level contrastive alignment objective that enhances the correspondence between visual segments and language representations.
- We conduct ablations and experiments on PHOENIX14T, demonstrating that our method achieves translation performances exceeding prior state-of-the-art.

2. Related Work

Sign Language Understanding (SLU) spans a wide range of tasks aimed at interpreting signed communication from visual input. Isolated Sign Language Recognition (ISLR) focuses on classifying individual signs from short, pre-segmented video clips [1, 14, 17, 21, 38, 48], while Continuous Sign Language Recognition (CSLR) targets the recognition of co-articulated sign sequences within continuous video streams. While both tasks are crucial in SLU, CSLR better reflects the natural temporal and linguistic structure of sign language and has been advanced considerably by gloss-annotated datasets such as PHOENIX14 [18] and CSL-Daily [46]. However, the lack of frame-level alignment in these datasets has led to the widespread adoption of Connectionist Temporal Classification (CTC) as a weakly supervised learning approach [15, 27, 28, 33, 47].

Despite the progress in CSLR, the broader objective of SLU lies in Sign Language Translation (SLT), the task of generating spoken-language translations directly from sign language videos. Unlike CSLR, which outputs gloss sequences, SLT aims to produce coherent natural language sentences, thus offering a more accessible communication channel for both Deaf and hearing communities. Camgoz et al. [3] first formalized SLT as a sequence-to-sequence learning problem, establishing a foundation for subsequent advancements. SLT methods can broadly be categorized into *gloss-based* and *gloss-free* approaches.

Gloss-based approaches rely on gloss annotations either as an explicit intermediate representation or as auxiliary supervision. For example, SLRT [4] employed a Transformer-based encoder-decoder architecture conditioned on gloss sequences. SignBT [46] improved on this pipeline by using back-translation, a concept adapted from machine translation. A multi-modal transfer learning-based (MMTLB) approach [6] was then proposed by pretraining on general-domain datasets, while CV-SLT [44] enhanced this by integrating a conditional variational autoencoder to improve cross-modal generalization. Other efforts, such as TS-SLT [7], also explored two-stream architectures that leverage both raw video and keypoint sequences.

In general, gloss-based methods significantly outperform gloss-free approaches due to the finer-grained supervision; however, they face the critical limitation of being dependent

on gloss annotations. This requirement poses a scalability challenge, as gloss annotations are labor-intensive and not consistently available across datasets and sign languages.

Gloss-free methods, in contrast, bypass gloss annotations entirely and learn directly from video and spoken language pairs. While they typically underperform gloss-based systems due to weaker supervision, they offer a scalable alternative by removing the need for gloss annotations. This approach is becoming increasingly viable with the emergence of large-scale gloss-free datasets such as YTSL-25 [36] and CSL-News [22]. As a result, the field has rapidly pivoted towards gloss-free SLT, prioritizing architecture designs and training strategies that can compensate for the absence of gloss through alternative learning signals.

For instance, GASLT [40] introduced a gloss attention mechanism to inject gloss-like inductive bias into the learning process. GFSLT [45], leveraged a CLIP-based [34] architecture, and was pretrained on video-sentence pairs via cross-modal alignment. Sign2GPT [37] introduced pseudo-gloss generation during pretraining to recover some level of intermediate supervision. SignCL [39] proposed a contrastive learning framework by treating temporally adjacent frames as semantically coherent units. FLa-LLM [8] incorporated a lightweight Transformer, and proposed pretraining with a language modeling objective. Lastly, SignLLM [12] explored a discrete codebook to learn gloss-like token representations in a self-supervised manner.

While all aforementioned methods showcase a wide range of strategies, most adopt Transformer-based encoders to process video frames, resulting in increased computation. In this work, we thus address these limitations by proposing a more efficient gloss-free SLT framework that maintains strong performances while reducing complexity.

3. Methodology

Our proposed framework for SLT addresses two key limitations in existing approaches: the use of linguistically-agnostic temporal sampling, which fails to respect the semantic boundaries of signs, and the high memory complexity of current Transformer-based models, which hinders scalability to larger datasets. To this end, we introduce a two-stage SLT architecture (in Fig. 2), that uses a novel visual tokenizer to segment continuous sign videos into discrete units, allowing for more memory efficient translation.

3.1. Stage 1: Visual-Language Pretraining

In the first stage, our goal is to pretrain a visual encoder such that it maps segmented sign units into semantically meaningful embeddings aligned with word-level linguistic representations. Instead of aligning directly to spoken sentences, which may contain syntactic noise, we adopt a pseudo-gloss formulation that provides a more compact and semantically focused target for alignment. We discuss our overall pretraining architecture further in the following subsections, including the segment-aware visual tokenization, pseudo-gloss extraction pipeline and contrastive learning objective.

3.1.1. Segment-Aware Visual Tokenization

Given a continuous sign language video consisting of T frames, denoted as $\mathbf{X} = \{x_1, x_2, \ldots, x_T\}$, our goal is to segment the video into N temporally localized sign units using motion and appearance cues. This results in a set of segments $\mathbf{X}' = \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_N\}$, where each $\mathbf{S}_i \in \mathbb{R}^{n_i \times c}$ corresponds to a contiguous sequence of n_i frames with c-dimensional image features. These segments are subsequently passed into a visual encoder (Section 3.1.3) to produce a sequence of sign tokens $\mathbf{t} = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_N\}$, where each $\mathbf{t}_i \in \mathbb{R}^c$ captures the motion dynamics and fine-grained spatial information of a localized sign unit.

To perform sign segmentation, we utilize the recently proposed *Hands-On* model [25], which segments signing sequences based on input features from HaMeR hand representations [30] and 3D body pose trajectories [16]. In our pipeline, the Hands-On model is used in a frozen state, without task-specific fine-tuning, relying on its learned priors over hand shape and upper-body motion to generalize across datasets. For additional details on the segmentation model, runtime analysis, and visual tokenization implementation, please refer to the supplementary materials.

3.1.2. Language Encoder

Pseudo-Gloss Extraction: Before extracting textual embeddings, we first convert the spoken sentence into a pseudo-gloss sequence that serves as a compact and semantically grounded alignment target. Given a spoken sentence $\mathbf{S} = \{w_1, w_2, \ldots, w_L\}$ of length L, we derive a pseudo-gloss sequence $\mathbf{G} = \{g_1, g_2, \ldots, g_M\}$, where $M < L$, by filtering for content-bearing words. Following [37], we apply a part-of-speech (POS) tagging pipeline and retain only tokens tagged as nouns, verbs, adjectives, numerals, adverbs, pronouns, or proper nouns. This process removes syntactic noise while preserving the semantic core of the sentence, yielding a gloss-like abstraction that is better aligned with the compositional structure of signing.

Textual Representation. To obtain linguistic features for alignment, we use *mBART-large-50* [35] as our language encoder. Thus, given a sequence of pseudo-glosses $\mathbf{G}$, the model first tokenizes them into a sequence of subword tokens $\{t_1, t_2, \ldots, t_M\}$. Each token t_i is then embedded via an input embedding matrix $E \in \mathbb{R}^{d \times 1024}$, with d as the vocabulary size, yielding:

$$\mathbf{T}^E = \{\mathbf{e}_i = E[t_i] \in \mathbb{R}^{1024} \mid i = 1, \ldots, M\}, \quad (1)$$

where token-level embeddings $\mathbf{T}^E$ capture lexical identity but lack contextual semantics, and serve as type-level lin-

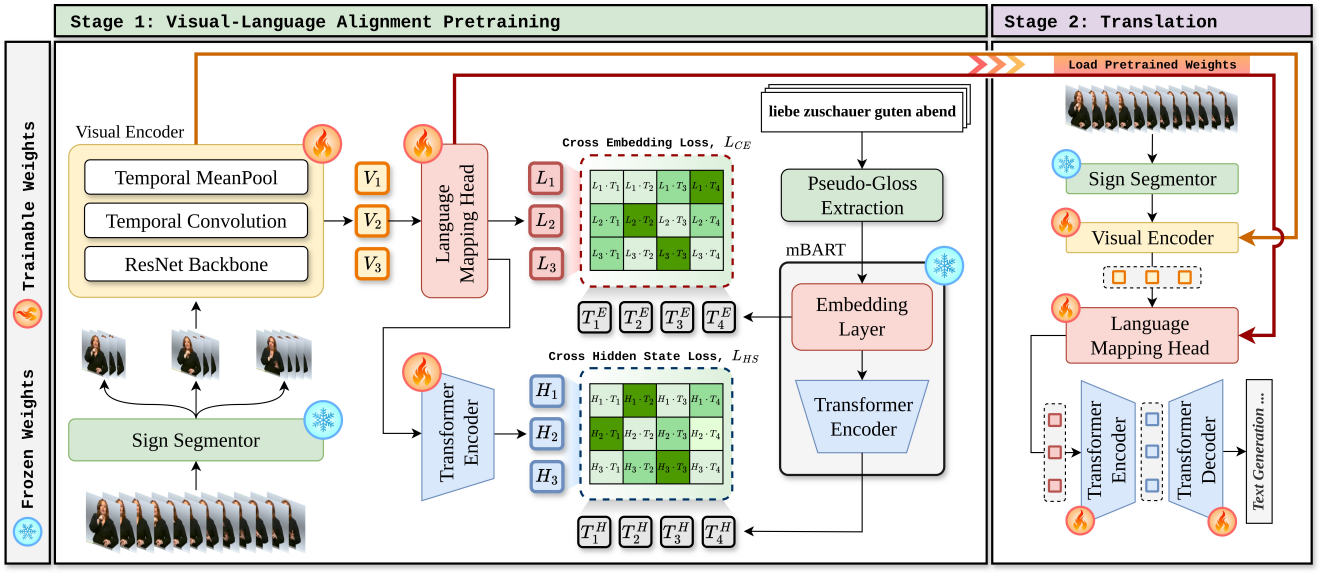

Figure 2. **Architecture Overview.** Our method consists of two stages. (i) In the first stage, a sign segmentor tokenizes continuous sign videos, and a visual encoder maps each segment into an embedding. We then apply contrastive pretraining to align the visual representations with the textual embeddings from a language encoder, encouraging semantically rich and discriminative features. (ii) In the second stage, the visual tokenizer and pretrained encoder are integrated into a sequence-to-sequence model and fine-tuned for sign language translation.

guistic representations in our contrastive alignment objective. To capture contextual dependencies between tokens, we pass $\mathbf{T}^E$ through the mBART Encoder ME to obtain contextualized hidden states:

$$\mathbf{T}^H = ME(\mathbf{T}^E) = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_M\}, \tag{2}$$

where each embedding $\mathbf{h}_i \in \mathbb{R}^{1024}$ encodes context-aware semantics via self-attention over the entire input sequence.

Unlike prior approaches that align visual features solely with the final hidden state outputs of a language encoder, we leverage both $\mathbf{T}^E$ and $\mathbf{T}^H$ in our contrastive objective. This dual representation enables supervision at both lexical and contextual levels, capturing stronger fine-grained and compositional structure inherent to sign language.

3.1.3. Visual Encoder

To represent the visual modality at the sign segment level, we construct a hierarchical encoder that captures both spatial and temporal structure across consecutive video frames. Each sign segment $S = \{s_1, s_2, \ldots, s_n\}$, comprising of n frames, is first processed by a ResNet-34 to extract frame-wise spatial features. This results in a sequence of embeddings $V = \{v_1, v_2, \ldots, v_n\}$, where $v_i \in \mathbb{R}^{512}$. Meanwhile, local temporal dependencies within the segment are captured via a lightweight temporal encoder consisting of a 1D convolutional layer (kernel size $k = 5$), followed by Batch Normalization and ReLU activation. In parallel, the temporal encoder also up-projects the features, resulting in a spatiotemporal representation of $V^{\text{temp}} \in \mathbb{R}^{(n-4)\times 1024}$.

We then apply average pooling across the temporal dimension to obtain a singular token representation $V^{\text{tok}} \in \mathbb{R}^{1024}$ that summarizes both spatial and temporal dynamics of the sign segment. Finally, to capture contextual relationships across segments, we introduce a Transformer-based encoder VE that operates over the sequence of tokens and outputs hidden states $\mathbf{H} \in \mathbb{R}^{1024}$. Since the token sequence is much shorter than the original frame sequence, our design enables efficient global reasoning while maintaining fine-grained semantic structure on the full signing sequence.

3.1.4. Vision-to-Language Mapper

To bridge the gap between visual and linguistic modalities, we introduce a *Visual-to-Language Mapper* that projects visual sign embeddings into the token embedding space of a pretrained language model. Given a visual token representation $V^{\text{tok}} \in \mathbb{R}^{1024}$, the mapper f_{map} produces a corresponding language embedding $\mathbf{L} = f_{\text{map}}(V^{\text{tok}}) \in \mathbb{R}^{1024}$. The function f_{map} is implemented as a stack of three feed-forward blocks, each consisting of Layer Normalization, a linear projection, GeLU activation, and Dropout, following Transformer-inspired architectures [10]. We supervise the mapper to align $\mathbf{L}$ with the token embeddings $\mathbf{T}^E$ from the language encoder, ensuring the mapped representations reside in a semantically meaningful space.

3.1.5. Contrastive Learning Objective

To enable fine-grained alignment between visual and textual tokens during pretraining, we require an objective that captures cross-modal correspondences without relying on strict

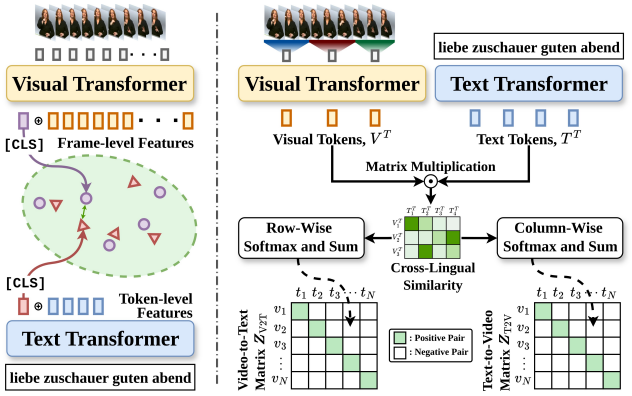

Figure 3. **Contrastive Learning Comparison.** Prior pretraining methods adopt CLIP-style objectives to align video and text globally via summarized [CLS] tokens. In contrast, our approach performs fine-grained cross-lingual alignment by computing token-level similarities across all the video-text pairs within a batch.

ordering. As our pseudo-glosses follow spoken language order rather than the temporal progression of the video, monotonic alignment strategies such as CTC are unsuitable.

Given these limitations, contrastive learning offers a more flexible alternative for cross-modal supervision. However, widely adopted CLIP-style methods [34] typically operate at the global level, compressing entire videos and sentences into single [CLS] embeddings. While proven effective for SLT [45], such global representations discard the rich internal structure of both modalities, making them less suited for the token-level alignment our framework requires.

To address this, we adopt a more granular supervision objective based on the *Cross-Lingual Contrastive Learning (CLCL)* loss [9], originally introduced to learn a joint sign video and spoken-language embedding space while capturing fine-grained sign-to-word mappings. Unlike the original implementation, which uses per-frame I3D features as sign representations, we apply CLCL at the token level by aligning segment-level visual tokens to language tokens.

In addition to token-level alignment, we also follow CLCL by computing a global video-to-text similarity matrix $\boldsymbol{Z}_{\text{V2T}} \in \mathbb{R}^{N\times N}$ for each mini-batch, where $\boldsymbol{Z}_{\text{V2T}}^{(i,j)}$ denotes the similarity between the i-th video and the j-th sentence. We then apply the InfoNCE loss [13], formulated as:

$$\mathcal{L}_{\text{V2T}} = -\frac{1}{2N}\sum_{i=1}^{N}\log\frac{\exp(\boldsymbol{Z}_{\text{V2T}}^{(i,i)}/\tau)}{\sum_{j=1}^{N}\exp(\boldsymbol{Z}_{\text{V2T}}^{(i,j)}/\tau)} - \frac{1}{2N}\sum_{j=1}^{N}\log\frac{\exp(\boldsymbol{Z}_{\text{V2T}}^{(j,j)}/\tau)}{\sum_{i=1}^{N}\exp(\boldsymbol{Z}_{\text{V2T}}^{(i,j)}/\tau)}, \tag{3}$$

where τ is a learnable temperature parameter. We similarly compute the text-to-video loss $\mathcal{L}_{\text{T2V}}$ by using the similarity matrix $\boldsymbol{Z}_{\text{T2V}}$, and define the final loss as a weighted sum:

$$\mathcal{L}_{\text{CLCL}} = \alpha\mathcal{L}_{\text{V2T}} + (1-\alpha)\mathcal{L}_{\text{T2V}}, \tag{4}$$

where $\alpha \in [0,1]$ controls the contribution of each alignment direction. Furthermore, as described in Section 3.1.2, we enhance the contrastive supervision by applying the CLCL loss to both the input embeddings $\mathbf{T}^E$ and the contextualized hidden states $\mathbf{T}^H$ of the language encoder. As the two contrastive objectives are applied separately we refer to them as the *Cross-Embedding Loss* $\mathcal{L}_{\text{CE}}$ and *Cross-Hidden-State Loss* $\mathcal{L}_{\text{HS}}$, respectively, as illustrated in Fig. 2. The final pretraining objective is then formulated as a weighted sum of the two, controlled by a hyperparameter $\beta \in [0,1]$ which is tuned through ablation experiments in Table 4b:

$$\mathcal{L}_{\text{total}} = \beta\,\mathcal{L}_{\text{CE}} + (1-\beta)\,\mathcal{L}_{\text{HS}}, \tag{5}$$

3.2. Stage 2: Sign Language Translation

In the second stage, our objective is to fine-tune our model for the downstream task of SLT. We begin by initializing the visual encoder and the Visual-to-Language Mapper with pretrained weights from the alignment stage (see Fig. 2). Since this stage already aligns visual embeddings with linguistic representations, it provides a strong language-aware prior for downstream translation. To complete the architecture, we attach the full *mBART-large-50* encoder-decoder model, where the mBART encoder consumes the mapped visual tokens and the decoder autoregressively generates the target spoken sentence. This setup allows for end-to-end translation of sign videos into spoken language sentences.

To ensure consistency with the pretraining pipeline and to maintain computational efficiency, we also retain the segment-based tokenizer to produce compact visual tokens. Compared to prior frame-based methods, this approach significantly reduces sequence length while preserving semantic granularity in SLT. Thus, given an input sign sequence $\mathbf{X} = \{x_1, x_2, \ldots, x_T\}$ of T frames, we first obtain a sequence of visual tokens $\mathbf{V} = \{v_1, \ldots, v_N\} \in \mathbb{R}^{N\times 1024}$ via the visual encoder. These are then projected into the linguistic space through the mapper, yielding $\mathbf{L} = \{l_1, \ldots, l_N\} \in \mathbb{R}^{N\times 1024}$, which are then processed by the mBART encoder to produce hidden states $\mathbf{H}$. The decoder subsequently generates the translated output sentence $\mathbf{S}' = \{s'_1, \ldots, s'_M\}$ token-by-token, and the entire model is optimized using the standard language modeling objective:

$$\mathcal{L}_{\text{LM}} = -\sum_{m=1}^{M}\log P(s'_m \mid s'_{<m}, \mathbf{X}), \tag{6}$$

where s'_m denotes the m-th predicted token and $s'_{<m}$ its preceding tokens. This objective thus trains the decoder to generate coherent translations conditioned on the visual input.

4. Experiments

4.1. Datasets and Evaluation Metrics

Dataset. To ensure comparability with prior work, we evaluate our model on the widely adopted *RWTH-PHOENIX-Weather 2014T (PHOENIX14T)* benchmark [3]. This dataset consists of German Sign Language (DGS) videos extracted from weather forecast broadcasts, and is annotated with both gloss-level transcriptions as well as corresponding spoken German sentences. It consists of 8,257 video clips, amounting to approximately 11 hours of footage. The standard splits include 7,096 samples for training, 519 for validation, and 642 for testing.

Evaluation Metrics. Following established SLT evaluation protocols, we report performance using *Bilingual Evaluation Understudy (BLEU)* [29] and *Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L)* [24]. BLEU measures the n-gram overlap between predicted and reference translations, capturing precision over different granularities. We report BLEU scores from unigram to 4-gram (BLEU-1 to BLEU-4). ROUGE-L evaluates translation quality based on the longest common subsequence, and considers both precision and recall. We follow the standard practice of reporting the F1-score variant of ROUGE-L.

4.2. Implementation Details

Stage 1: Contrastive Pretraining. We train the visual encoder and the Vision-to-Language Mapper using a batch size of 16 and stochastic gradient descent (SGD) with an initial learning rate of 0.03, momentum of 0.9, and a cosine annealing learning rate schedule. Gradient clipping is applied with a threshold of 1.0, and a dropout rate of 0.1 is used in the mapper to mitigate overfitting. Training is performed for 80 epochs on a single NVIDIA A100 GPU.

To enhance generalization, we apply slight data augmentations following [45], including random rotation, color jitter, and brightness adjustments. All video frames are resized to 256×256, followed by center cropping to 224×224 to ensure input consistency. In line with prior work, we also trim the mBART tokenizer's embedding matrix to retain only vocabulary tokens present in the training set.

Stage 2: Downstream Translation. When fine-tuning for SLT, we use a batch size of 8 and optimize with SGD. The learning rate is set to 0.004, momentum to 0.9, and weight decay to 0.001. Gradient clipping is set to 5.0, and dropout remains at 0.1. Training is again conducted for 80 epochs on a single NVIDIA A100. Similar data augmentations are applied and the model is trained using cross-entropy with label smoothing of 0.2. During inference, we configure the mBART decoder with a maximum token length of 150 and set number of beams to 4 for decoding.

Method	B1	B2	B3	B4	R-L
Gloss-Based					
SL-T [4]	46.61	33.73	26.19	21.32	–
SignBT [46]	50.80	37.75	29.72	24.32	49.54
MMTLB [6]	53.97	41.75	33.84	28.39	52.65
SLTUNET [42]	52.92	41.76	33.99	28.47	52.11
TS-SLT [7]	54.90	42.43	34.46	28.95	53.48
CV-SLT [44]	54.88	42.68	34.79	29.27	54.33
Gloss-Free					
NSLT [3]	29.86	17.52	11.96	9.00	30.70
TSPNet [20]	36.10	23.12	16.88	13.41	34.96
CSGCR [43]	36.71	25.40	18.86	15.18	38.85
GASLT [40]	39.07	26.74	21.86	15.74	39.86
GFSLT-VLP [45]	43.71	33.18	26.11	21.44	42.49
Sign2GPT [37]	49.54	35.96	28.83	22.52	**48.90**
FLa-LLM [8]	46.29	35.33	28.03	23.09	45.27
SignLLM [12]	45.21	34.78	28.05	23.40	44.49
SAGE (Ours)	**49.69**	**37.01**	**29.21**	**24.10**	48.86

Table 1. **Sign Language Translation results on PHOENIX14T (test set).** We compare against recent SOTA gloss-free SLT models. **Bold** indicates best performance among gloss-free methods.

4.3. Comparisons to State-of-the-Art Methods

To assess our token-efficient approach, we conduct extensive comparisons against recent state-of-the-art (SOTA) methods. We evaluate both the maximum translation performance (Section 4.3.1) and the computational efficiency (Section 4.3.2), including comparisons of total memory usage and performance under similar token reduction settings.

4.3.1. Maximum Translation Performance Comparisons

In Table 1, we compare our model against recent state-of-the-art (SOTA) methods on the PHOENIX14T test set. Our approach achieves the highest performance across all BLEU scores, surpassing previous methods, including a +0.7 increase in BLEU-4 (24.10 vs. 23.40). The gains in BLEU-1 to BLEU-3 suggest that our token-level supervision is particularly effective at capturing fine-grained, word-level correspondences. This is likely due to our design, where sign videos are represented as discrete tokens and directly aligned with text tokens during training, successfully reducing the gap between visual input and linguistic output at the word level. Beyond token-level matches, our improvement in BLEU-4 further indicates strong performance at the sentence level. This suggests that the visual tokens not only provide accurate local alignments, but also serve as robust priors for the language decoder to generate coherent and fluent translations. Additionally, our ROUGE score closely matches that of Sign2GPT, highlighting that our model preserves strong semantic similarity to the reference translations. Importantly, we achieve these results with a signifi-

Method	MP	SD	SW	UD	Ratio	B4
GFSLT† [45]	✓	✗	✗	–	0.25	21.44
	✓	✗	✗	T/2	<u>0.125</u>	<u>19.25</u>
Sign2GPT [37]	✗	✓	✗	T/2	0.25	22.52
FLa-LLM [8]	✗	✗	✗	T/4	0.25	23.09
	✗	✗	✗	T/8	<u>0.125</u>	<u>20.02</u>
SignLLM [12]	✗	✗	✓	–	0.25	23.40
SAGE (Ours)	**Visual Tokenization**				**0.129**	**24.10**

Table 2. **Comparison of Temporal Reduction Strategies.** We compare our Visual Tokenization approach with Max Pooling (MP), Strided Downsampling (SD), Sliding Window (SW), and Uniform Downsampling (UD). Reported reduction ratios provide context on computational efficiency, with those comparable to ours <u>underlined</u> for emphasis. † indicates results reproduced by us.

Method	VRAM (Total)	Base GPU Req.
GFSLT [45]	80 GB	1 × A100
Sign2GPT [37]	160 GB	2 × A100
SignLLM [12]	96 GB	4 × A5000
SAGE (Ours)	**∼ 60 GB**	**∼ 3 × RTX 3090**

Table 3. **Peak GPU Memory Usage During Training.** We report the peak VRAM consumption and minimum GPU required based on the training setups used by each method to achieve their respective reported performances. For our method, the reported values are approximate due to minor variation in input sequence lengths.

cantly more efficient and lightweight architecture, demonstrating the potential of visual tokenization as an effective and scalable representation for sign language translation.

4.3.2. Computational Efficiency Comparisons

To further contextualize our contribution, we evaluate how prior SOTA models perform under comparable sequence reduction constraints. Existing approaches typically reduce input frame lengths using fixed downsampling strategies, such as strided sampling, sliding windows, or pooling; and they are often applied at a factor of 4 (i.e., a reduction ratio of 0.25). In contrast, our method employs a visual tokenization strategy that maps input frames into a compact sequence of semantic tokens, achieving a significantly higher compression rate with a reduction ratio of 0.129, nearly halving the sequence length compared to typical methods.

To assess the impact of such compression on translation quality, we evaluate how prior models perform under similar levels of temporal reduction. Specifically, we retrain GFSLT using T/2 frame downsampling and include the reported T/8 results for FLa-LLM [8] from their original work. While such downsampling strategies are commonly used, we acknowledge that they may substantially degrade the fidelity of the input signal. However, these strategies represent some of the only practical options currently available in SLT for reducing sequence length, highlighting the need for more semantically informed compression strategies. As shown in Table 2, these models exhibit notable performance drops under increased compression; with FLa-LLM, for instance, falling to a BLEU-4 of 20.02 (-3.07). In contrast, our approach offers strong translation performance (BLEU-4: 24.10) despite the compact sequence length.

Table 3 compares the peak GPU memory consumption of our model against recent SOTA. This comparison is especially important given the reliance of SLT architectures on contrastive learning, which is highly sensitive to the *effective batch size* during training. Larger batch sizes allow for a greater number of negative pairs; however, achieving such large batches is often constrained by GPU memory. Fortunately, our method requires significantly less peak GPU memory compared to prior SOTA, using up to 2.67× less VRAM than Sign2GPT, resulting in more efficient training and improved scalability to larger-scale datasets.

4.4. Ablation Study

To assess the effectiveness of various architectural and training design choices in our model, we conduct extensive ablation studies on the PHOENIX14T development set, focusing on the BLEU-4 and ROUGE-L scores.

Pretraining Loss Configuration. We first investigate the effect of different contrastive loss formulations and the role of our proposed dual-level supervision. As shown in Table 4a, using the fine-grained CLCL loss consistently outperforms CLIP-style global alignment, demonstrating the importance of token-level contrastive supervision in closing the visual-language gap. Furthermore, results also indicate that dual-level supervision within the language encoder provides a stronger alignment signal, as it enhanced model performances regardless of whether CLIP or CLCL was used.

Loss Balancing. In Table 4b, we then ablate the weighting coefficient β between the Cross-Embedding and Cross-Hidden-State losses as defined in Eq. 5. Varying β from 0 to 0.8, we find that $\beta = 0.6$, which places more emphasis on the embedding-level loss, achieves the highest BLEU-4 score, while maintaining a ROUGE-L score nearly on par with the best. These results suggest that prioritizing alignment at the earlier language representation stage provides a stronger supervisory signal for downstream translation.

Pretraining Paradigm. Finally, we examine how different pretraining configurations affect downstream translation performance, as shown in Table 5. Training the model end-to-end without any pretraining yields significantly lower performance (BLEU-4: 16.01, ROUGE-L: 39.23), highlighting the importance of visual-language alignment. In contrast, introducing a lightweight pretraining stage, where

Loss	DS	BLEU4	ROUGE
CLIP	✗	21.93	47.04
CLIP	✓	22.30	47.88
CLCL	✗	23.25	48.10
CLCL	✓	**23.81**	**49.08**

(a) Different Loss Configurations

β	BLEU4	ROUGE
0	23.25	48.10
0.2	23.58	**49.11**
0.4	23.25	48.57
0.6	**23.81**	49.08
0.8	22.78	48.54

(b) Different β values

Table 4. **Ablation Study on Pretraining Loss Design.** We present two sets of ablations: (a) a comparison of different loss configurations, evaluating CLIP-style versus CLCL supervision, and the impact of incorporating dual-level supervision (DS); and (b) the effect of hyperparameter β in the final loss formulation (Eq. 5).

Pretrained Setup	Init in Stage 2	BLEU4	ROUGE
None	None	16.01	39.23
VLE	VLE	21.70	46.77
VLE + 3L TE	**VLE**	**23.81**	**49.08**
VLE + 12L TE	VLE	23.43	48.67
VLE + 12L TE	VLE + TE	22.88	48.38

Table 5. **Ablation of Pretraining Configurations.** We compare different encoder setups (**VLE**: Visual-Language Encoder, **TE**: Transformer Encoder with varying layers L) in pretraining and weight loading strategies for downstream translation.

only the Visual Encoder and Language Mapper are trained using the Cross-Embedding Loss, leads to a notable improvement (BLEU-4: 21.70, ROUGE-L: 46.77), demonstrating that even shallow alignment provides strong benefits. Further adding a Transformer module during pretraining then enables dual supervision via both Cross-Embedding and Cross-Hidden-State losses. Interestingly, the 3-layer Transformer configuration outperforms its 12-layer counterpart, suggesting that deeper models may introduce unnecessary complexity and hinder alignment. Surprisingly, the best results are achieved when the 3-layer Transformer is used during pretraining but discarded during fine-tuning (BLEU-4: 23.81, ROUGE-L: 49.08). This highlights that while lightweight contextualization improves alignment, removing it downstream may enhance transferability by reducing overfitting to pretraining objectives.

4.5. Qualitative Results

Fig. 4 shows the similarity matrix between segmented visual tokens and pseudo-gloss candidates via dot-product of their embeddings. The clear diagonal patterns show that the CLCL objective effectively enforces fine-grained, token-level alignment with semantically relevant pseudo-glosses.

However, since pseudo-glosses are derived from spoken language, they may not always accurately reflect the content of the signed video. For example, in Fig. 4, terms such as *Breiten*, *Teilweise*, and *Ergiebig* are present in the pseudo-gloss despite not being signed, while actual glosses

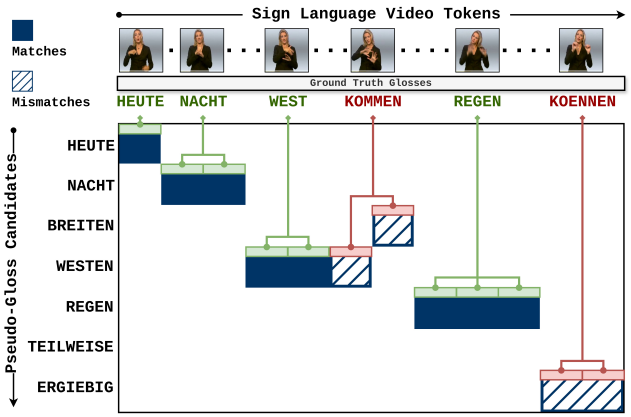

Figure 4. **Visual-to-Text Token Similarity Matrix.** The CLCL objective enables accurate fine-grained alignment between visual tokens and pseudo-gloss candidates. The solid blue boxes indicate correct matches, while the striped blue boxes mark mismatches.

like *Kommen* and *Koennen* are omitted. Such discrepancies can lead to occasional misalignments, as visual tokens may be left unmatched when no valid pseudo-gloss is present.

Despite these limitations, the pseudo-gloss representation still captures the majority of relevant glosses and facilitates effective alignment without requiring manually annotated gloss labels. This highlights the effectiveness of our framework, where segmentation-based tokenization and the CLCL objective jointly enables the learning of semantically meaningful alignments even under weak supervision. For further qualitative results, including additional alignment maps, spoken language predictions, and segment-level visualizations, please refer to the supplementary material.

5. Conclusions

In this work, we introduced a segment-aware tokenization framework for gloss-free SLT, leveraging sign segmentation for both localization and temporal downsampling. This reduces sequence length by 2× and memory usage by 2.67×, enabling efficient and scalable training. We further proposed a dual-level alignment loss that enhances cross-modal supervision at both the input and language representation levels. On PHOENIX14T, our method surpasses prior state-of-the-art, especially under matched sequence lengths. These results highlight the effectiveness of our approach for efficient gloss-free SLT. Future work will focus on improving the visual encoders and scaling to larger corpora.

Acknowledgments: This work was supported by the SNSF project 'SMILE II' (CRSII5 193686), the Innosuisse IICT Flagship (PFFS-21-47), EPSRC grant APP24554 (SignGPT) and through funding from Google.org via the AI for Global Goals scheme. This work reflects only the author's views and the funders are not responsible for any use that may be made of the information it contains.

References

[1] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In *ECCV*, pages 35–53. Springer, 2020. 2

[2] Carl Börstell, Thomas Hörberg, and Robert Östling. Distribution and duration of signs and parts of speech in swedish sign language. *Sign Language & Linguistics*, 19(2):143–196, 2016. 2

[3] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *CVPR*, pages 7784–7793, 2018. 1, 2, 6

[4] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *CVPR*, pages 10023–10033, 2020. 1, 2, 6

[5] Amanda Cardoso Duarte, Shruti Palaskar, Lucas Ventura Ripol, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres Viñals, and Xavier Giró Nieto. How2sign: A large-scale multimodal dataset for continuous american sign language. In *CVPR*, pages 2734–2743. Institute of Electrical and Electronics Engineers (IEEE), 2021. 1

[6] Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. A simple multi-modality transfer learning baseline for sign language translation. In *CVPR*, pages 5120–5130, 2022. 1, 2, 6

[7] Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. Two-stream network for sign language recognition and translation. *NeurIPS*, 35:17043–17056, 2022. 2, 6

[8] Zhigang Chen, Benjia Zhou, Jun Li, Jun Wan, Zhen Lei, Ning Jiang, Quan Lu, and Guoqing Zhao. 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)*, pages 7071–7081, 2024. 2, 3, 6, 7

[9] Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *CVPR*, pages 19016–19026, 2023. 5

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019. 4

[11] Edward Fish and Richard Bowden. Geo-sign: Hyperbolic contrastive regularisation for geometrically aware sign language translation. *arXiv preprint arXiv:2506.00129*, 2025. 2

[12] Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. Llms are good sign language translators. In *CVPR*, pages 18362–18372, 2024. 1, 2, 3, 6, 7

[13] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 5

[14] Hezhen Hu, Wengang Zhou, and Houqiang Li. Hand-model-aware sign language recognition. In *AAAI*, pages 1558–1566, 2021. 2

[15] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Continuous sign language recognition with correlation network. In *CVPR*, pages 2529–2539, 2023. 2

[16] Maksym Ivashechkin, Oscar Mendez, and Richard Bowden. Improving 3d pose estimation for sign language. In *ICASSP*, pages 1–5. IEEE, 2023. 3

[17] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Skeleton aware multi-modal sign language recognition. In *CVPR*, pages 3413–3423, 2021. 2

[18] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, 2015. 2

[19] Oscar Koller, Necati Cihan Camgoz, Hermann Ney, and Richard Bowden. Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *IEEE TPAMI*, 42(9):2306–2320, 2019. 1

[20] Dongxu Li, Chenchen Xu, Xin Yu, Kaihao Zhang, Benjamin Swift, Hanna Suominen, and Hongdong Li. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *NeurIPS*, 33:12034–12045, 2020. 6

[21] Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, and Hongdong Li. Transferring cross-domain knowledge for video sign language recognition. In *CVPR*, pages 6205–6214, 2020. 2

[22] Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. Uni-sign: Toward unified sign language understanding at scale. *arXiv preprint arXiv:2501.15187*, 2025. 2, 3

[23] Diane C Lillo-Martin and Jon Gajewski. One grammar or two? sign languages and the nature of human language. *Wiley Interdisciplinary Reviews. Cognitive Science*, 5(4): 387–401, 2014. 1

[24] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004. 6

[25] Jianhe Low, Harry Walsh, Ozge Mercanoglu Sincan, and Richard Bowden. Hands-on: Segmenting individual signs from continuous sequences. In *FG*, 2025. 2, 3

[26] Jiachen Lu, Jinghan Yao, Junge Zhang, Xiatian Zhu, Hang Xu, Weiguo Gao, Chunjing Xu, Tao Xiang, and Li Zhang. Soft: Softmax-free transformer with linear complexity. *NeurIPS*, 34:21297–21309, 2021. 2

[27] Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen. Visual alignment constraint for continuous sign language recognition. In *ICCV*, pages 11542–11551, 2021. 2

[28] Zhe Niu and Brian Mak. Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. In *ECCV*, pages 172–186. Springer, 2020. 2

[29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[30] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *CVPR*, pages 9826–9836, 2024. 3

[31] Harvey P Peet. Elements of the language of signs. *American Annals of the Deaf and Dumb*, 5(2):83–95, 1853. 1

[32] Pamela Perniss, Roland Pfau, and Markus Steinbach. Can't you see the difference? sources of variation in sign language structure. *Emotion*, pages 1–34, 2007. 1

[33] Junfu Pu, Wengang Zhou, and Houqiang Li. Iterative alignment network for continuous sign language recognition. In *CVPR*, pages 4165–4174, 2019. 2

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 2, 3, 5

[35] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. 2020. 3

[36] Garrett Tanzer and Biao Zhang. Youtube-sl-25: A large-scale, open-domain multilingual sign language parallel corpus. *arXiv preprint arXiv:2407.11144*, 2024. 3

[37] Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. Sign2gpt: Leveraging large language models for gloss-free sign language translation. In *ICLR*. 1, 2, 3, 6, 7

[38] Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. Learnt contrastive concept embeddings for sign recognition. In *ICCV*, pages 1945–1954, 2023. 2

[39] Jinhui Ye, Xing Wang, Wenxiang Jiao, Junwei Liang, and Hui Xiong. Improving gloss-free sign language translation by reducing representation density. In *NeurIPS*, 2024. 2, 3

[40] Aoxiong Yin, Tianyun Zhong, Li Tang, Weike Jin, Tao Jin, and Zhou Zhao. Gloss attention for gloss-free sign language translation. In *CVPR*, pages 2551–2562, 2023. 3, 6

[41] Kayo Yin and Jesse Read. Better sign language translation with STMC-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics. 1

[42] Biao Zhang, Mathias Müller, and Rico Sennrich. Sltunet: A simple unified model for sign language translation. *arXiv preprint arXiv:2305.01778*, 2023. 1, 6

[43] Jian Zhao, Weizhen Qi, Wengang Zhou, Nan Duan, Ming Zhou, and Houqiang Li. Conditional sentence generation and cross-modal reranking for sign language translation. *IEEE TMM*, 24:2662–2672, 2022. 6

[44] Rui Zhao, Liang Zhang, Biao Fu, Cong Hu, Jinsong Su, and Yidong Chen. Conditional variational autoencoder for sign language translation with cross-modal alignment. In *AAAI*, pages 19643–19651, 2024. 2, 6

[45] Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. Gloss-free sign language translation: Improving from visual-language pretraining. In *ICCV*, pages 20871–20881, 2023. 1, 2, 3, 5, 6, 7

[46] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *CVPR*, pages 1316–1325, 2021. 1, 2, 6

[47] Ronglai Zuo and Brian Mak. C2slr: Consistency-enhanced continuous sign language recognition. In *CVPR*, pages 5131–5140, 2022. 2

[48] Ronglai Zuo, Fangyun Wei, and Brian Mak. Natural language-assisted sign language recognition. In *CVPR*, pages 14890–14900, 2023. 2