Contrastive Pretraining with Dual Visual Encoders for Gloss-Free Sign Language Translation

Ozge Mercanoglu Sincan
Centre for Vision, Speech and Signal Processing (CVSSP)
University of Surrey
Guildford, Surrey, United Kingdom
o.mercanoglusincan@surrey.ac.uk

Richard Bowden
Centre for Vision, Speech and Signal Processing (CVSSP)
University of Surrey
Guildford, Surrey, United Kingdom
r.bowden@surrey.ac.uk

Abstract

Sign Language Translation (SLT) aims to convert sign language videos into spoken or written text. While early systems relied on gloss annotations as an intermediate supervision, such annotations are costly to obtain and often fail to capture the full complexity of continuous signing. In this work, we propose a two-phase, dual visual encoder framework for gloss-free SLT, leveraging contrastive visual–language pretraining. During pretraining, our approach employs two complementary visual backbones whose outputs are jointly aligned with each other and with sentence-level text embeddings via a contrastive objective. During the downstream SLT task, we fuse the visual features and input them into an encoder–decoder model. On the Phoenix-2014T benchmark, our dual encoder architecture consistently outperforms its single-stream variants and achieves the highest BLEU-4 score among existing gloss-free SLT approaches.

CCS Concepts

• **Human-centered computing** → **Accessibility technologies**.

Keywords

Sign Language Translation, Gloss-free SLT, Contrastive Learning, Multimodal Pretraining

ACM Reference Format:
Ozge Mercanoglu Sincan and Richard Bowden. 2025. Contrastive Pretraining with Dual Visual Encoders for Gloss-Free Sign Language Translation. In *ACM International Conference on Intelligent Virtual Agents (IVA Adjunct '25), September 16–19, 2025, Berlin, Germany*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3742886.3756703

1 Introduction

Sign languages are visual languages that rely on both manual (handshape, orientation, movement) and non-manual (facial expressions, head and body posture) components. These components operate simultaneously and often encode grammatical information in parallel channels. With over 200 sign languages in the world [1], the development of inclusive technologies to support sign language understanding has become an increasingly important research domain.

Sign Language Translation (SLT) aims to convert sign language videos into spoken or written text. Achieving this requires the models to capture both fine-grained spatial detail (e.g., specific handshapes and finger configurations), temporal dynamics (e.g., motion trajectories). These visual features are then transformed into coherent natural language output. Works in this domain have predominantly relied on sequential models, such as Recurrent Neural Networks (RNNs) or the Transformer [33], to map sign language video sequences to spoken language sentences [3, 4, 30, 36, 38]. Earlier studies relied on intermediate annotations called glosses – simplified written descriptions of individual signs – to act as a bridge between visual and linguistic domains [3, 4, 38]. However, manual annotation of glosses is labor-intensive and difficult to annotate accurately, resulting in a limited number of publicly available gloss-annotated datasets [15].

To overcome this limitation, gloss-free SLT methods have been proposed, aiming to directly translate sign language videos into spoken language without relying on gloss supervision [8, 11, 37, 39]. This shift removes the dependence on expert gloss annotations but brings challenges in learning rich, discriminative representations of signing sequences. Typical architectures use a frozen visual encoder backbone, such as 3D Convolutional Neural Networks, e.g., I3D [5], to extract features, and pass them into transformers [2, 29, 30]. However, these models often lack a strong semantic bridge between visual and linguistic modalities, limiting their ability to generate semantically accurate translations.

Recent advancements in gloss-free SLT have drawn inspiration from visual-language pretraining methods [8, 34, 37, 39]. Approaches such as GFSLT-VLP [39] pioneered the use of contrastive learning to align visual representations with textual descriptions. These methods offer a promising direction to improve the semantic grounding of visual features and provide more flexible and scalable training paradigms for SLT.

In this work, we propose a dual visual encoder framework for gloss-free SLT. We employ two complementary visual backbones: a ResNet [13] to extract fine-grained spatial features from individual frames, and an I3D [5] to capture spatio-temporal features. During our pretraining stage, both feature sets are projected into a shared embedding space and aligned via a contrastive loss against corresponding sentence embeddings extracted by a pretrained language model. We additionally include an inter-modal alignment that encourages the ResNet and I3D features from the same video to be similar while pushing apart features from different video samples. After pretraining, we fine-tune a translation model for translation

IVA Adjunct '25, Berlin, Germany

ACM ISBN 979-8-4007-1996-7/25/09
https://doi.org/10.1145/3742886.3756703

by fusing visual features and feeding them into an mBART-based encoder-decoder architecture. We evaluate our approach on the Phoenix-2014T dataset [3] and demonstrate significant improvements over our single-encoder baselines. We outperform existing gloss-free SLT methods on the BLEU-4 metric, while obtaining competitive results on other evaluation metrics.

- We introduce a contrastively pretrained dual visual encoder architecture that jointly aligns both visual modalities with each other and with language. This alignment encourages the model to learn complementary representations and enhance semantic consistency.
- We explore and compare various fusion strategies for combining visual features.

2 Related Work

Recent research at the intersection of vision and language has led to rapid advances in many tasks, including sign language translation. In this section, we review works on visual-language pretraining and sign language translation.

2.1 Visual-Language Pretraining

Contrastive pretraining between vision and language has become fundamental for many downstream tasks. A pioneering work in this field, CLIP (Contrastive Language-Image Pre-training) [25] aligns image features with text representations using a contrastive learning strategy. CLIP does not need task-specific training data, and it demonstrates an ability to generalize across a wide range of downstream tasks such as zero-shot transfer to image classification, facial emotion recognition, optical character recognition, and so on. VideoCLIP [35] extends this contrastive learning objective to pre-train a unified video-text representation that surpasses prior works in video understanding tasks such as text-video retrieval, video question answering, and action segmentation.

Several extensions have adapted this idea for additional modalities. CLIP4VLA [26] extends CLIP by incorporating audio as a third modality, forming a unified tri-encoder structure for multimodal learning. IMAGEBIND [10] scales this idea further by binding six modalities (e.g., image, text, video, audio, depth, and thermal) – using images as a key modality, and aligning each modality to image embeddings. More recently, LANGUAGEBIND [41] proposes a language-based multi-modal pretraining, taking the language as the bind across video, infrared, depth, and audio.

2.2 Sign Language Translation

Sign language translation presents unique challenges compared to other vision-language tasks due to its complexity. Early studies [3, 4, 6, 7, 38] decomposed the SLT task two sub-tasks: Sign Language Recognition (SLR), recognizing sequences of signs, and then 'translating' the recognized signs into spoken language sentences. These gloss-based pipelines benefit from structured intermediate supervision, enabling models to focus on smaller vocabularies and more discrete targets. However, gloss annotations are highly labor-intensive and not available for large datasets, motivating gloss-free SLT methods.

Gloss-free SLT approaches aim to directly translate sign language videos into spoken language without gloss annotation. Initial gloss-free efforts (Sign2Text) typically adopted end-to-end encoder–decoder frameworks using a visual encoder and transformer-based decoders [4]. While these architectures removed the need for glosses, their performance was often inferior to gloss-based counterparts, largely due to weaker semantic grounding between visual features and linguistic output.

To address this, several works have focused on enhancing visual representations. Some approaches [28, 30] pretrain a classifier on sign language datasets and use it as a frozen visual encoder for the downstream SLT task. However, these approaches have a large semantic gap to solve between video features and text tokens. To address this, SignLLM [11] proposes a tokenization strategy, converting sign videos into a language-like format such as discrete character-level sign tokens and word-level sign representations. Then, they are passed to a large language model, LLaVA [17].

In parallel, some works have focused on visual-language pretraining methods. GFSLT-VLP [39] stands out as one of the first methods to leverage contrastive visual-language pretraining for gloss-free sign language translation, aligning visual features directly with textual representations. Building on this, SignCL [37] identified a representation density problem, where semantically distinct but visually similar signs tend to be close in the feature space, and proposed a contrastive learning strategy on adjacent frames to improve feature separation. Sign2GPT [34] proposes generating pseudo-glosses from spoken language descriptions during pretraining, offering an alternative to manual gloss annotations. Differently from these contrastive alignment methods, works such as FLa-LLM [8] and VAP [14] incorporate a lightweight translation model in the pretraining stage. VAP employed a multi-stage pretraining, including pseudo-gloss-to-text translation, which may be considered as weakly supervised gloss-free method.

Our approach draws inspiration from multimodal contrastive frameworks but focuses on visual–textual alignment in the context of sign language. We leverage the complementary strengths of two different visual encoders while aligning them both with each other and with language representations. This multiple-alignment strategy enables the model to capture complementary visual information from different aspects of the sign language video, leading to improved translation performance.

3 Method

We propose a two-stage framework, namely DVE-SLT, for gloss-free sign language translation. Given a sign video $V = (I_1, I_2, I_3, .., I_T)$ with T frames, the goal is to generate a spoken-language sentence $S = (w_1, w_2, .., w_U)$ with U words without relying on gloss annotations. Our architecture is illustrated in Fig. 1. In this section, we describe our pretraining strategy in which cross-modal and inter-modal alignment objectives are optimized through contrastive learning (Section 3.1), and the fine-tuning stage in which the fused visual features are decoded into spoken language (Section 3.2).

3.1 Visual-Language Pretraining

Dual Visual Encoders. Visual feature extraction is a critical component for video understanding tasks, including SLT. In prior works,

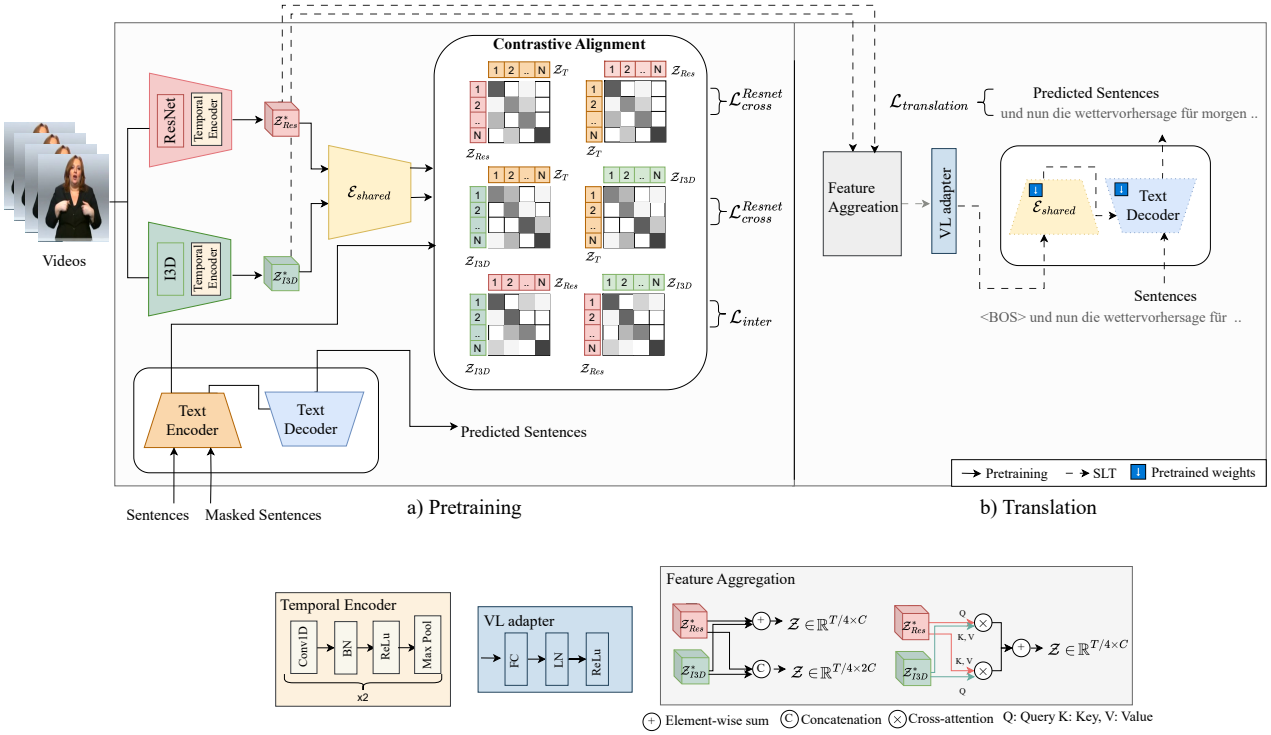

Figure 1: Overview of the proposed framework which has two phases: (a) pretraining, (b) translation. In pretraining, visual and textual representations are aligned utilizing contrastive alignment across modalities. The shared visual encoder, $\mathcal{E}_{shared}$, and Text Decoder weights are then utilized in the downstream SLT task. (c) Module structures, including the Temporal Encoder, VL Adapter, and Feature Aggregation are detailed.

a variety of visual encoders have been explored to extract meaningful representations. Among these, ResNet [13] and I3D [5] are frequently employed for sign language translation. ResNet-based models are commonly used for their efficiency and ability to capture fine-grained spatial features, by combining with temporal convolution layers to model temporal features [8, 37, 39]. On the other hand, I3D, originally developed for action recognition, is designed to capture spatio-temporal features simultaneously. It achieved notable success in sign language recognition and translation tasks [20, 28, 30].

In this work, we adopt both spatial (ResNet18) and spatio-temporal (I3D) visual encoders to extract complementary visual features. The visual encoders are followed by identical Temporal Encoders, which are composed of two blocks of a 1D-convolutional layer, a batch normalization layer, ReLu activation, and max pooling, reducing the number of frames to $T/4$. Although I3D inherently models temporal dynamics, we apply the same temporal module to both branches to ensure consistent temporal resolution. Projecting the outputs of both branches into a unified embedding space is critical for contrastive alignment. Therefore, we employ a single shared transformer encoder, $\mathcal{E}_{shared}$, which processes both ResNet and I3D features using the same parameters. This promotes semantic consistency between modalities and enables joint optimization during pretraining.

Text Encoder and Decoder. We employ the pretrained multilingual mBART [18] encoder with 12 layers to convert spoken language sentences into text embeddings, which are aligned with visual features via contrastive learning.

In parallel, masked versions of the same sentences are fed into the text encoder and then passed to an mBART decoder with 3 layers, which is trained to reconstruct the original sentence using a standard cross-entropy loss. During the translation stage, this decoder is used to initialize the language decoder of our SLT model, enabling a smooth transition from pretraining to fine-tuning.

Contrastive Alignment Objectives. During pretraining, we apply a *cross-modal* contrastive loss to align video and text embeddings. In addition, we introduce an *inter-modal* contrastive loss that aligns the representations from two complementary visual encoders for the same video. This dual-objective setup encourages both visual–text alignment and consistency across visual modalities, reinforcing the semantic coherence of the learned representations.

Given a batch of N video–text pairs, we extract visual features using our shared visual encoder and textual features using a pretrained mBART multilingual text encoder. Then, for each video-text pair, we compute a similarity matrix by adopting Cross-Lingual Contrastive Learning (CiCO) [9], which was proposed for the sign language retrieval task. We use a symmetric InfoNCE loss [12] to contrast matching and mismatching pairs:

$$\mathcal{L}_{\text{cross}} = -\frac{1}{2N}\sum_{i=1}^{N}\log\frac{\exp(Sim(Z_{v_i}, Z_{t_i})/\tau)}{\sum_{j=1}^{N}\exp(Sim(Z_{v_i}, Z_{t_j})/\tau)} - \frac{1}{2N}\sum_{j=1}^{N}\log\frac{\exp(Sim(Z_{v_j}, Z_{t_j})/\tau)}{\sum_{i=1}^{N}\exp(Sim(Z_{v_i}, Z_{t_j})/\tau)} \quad (1)$$

where $Sim(Z_{v_i}, Z_{t_j})$ denotes the global similarity of the i^{th} video embedding and the j^{th} text embedding, resulting the similarity matrix $\mathbf{M} \in \mathbb{R}^{N\times N}$ computed over the batch, τ is a trainable temperature parameter. The visual embedding Z_v is obtained from either the ResNet or I3D encoder.

In addition, since we employ two different visual encoders to extract complementary features from the same video, we introduce a second contrastive loss, *inter-modal contrastive loss*. It aims to pull together the embeddings from both branches for the same input, while pushing apart mismatched visual pairs. This objective encourages semantic alignment across encoders, reinforcing the consistency of dual visual representations. Similarly, we use the symmetric InfoNCE loss [12] but with the visual embeddings Z_{Res} and Z_{I3D}, extracted from the ResNet and I3D branches:

$$\mathcal{L}_{\text{inter}} = -\frac{1}{2N}\sum_{i=1}^{N}\log\frac{\exp(Sim(Z_{Res_i}, Z_{I3D_i})/\tau)}{\sum_{j=1}^{N}\exp(Sim(Z_{Res_i}, Z_{I3D_j})/\tau)} - \frac{1}{2N}\sum_{j=1}^{N}\log\frac{\exp(Sim(Z_{I3D_j}, Z_{Res_j})/\tau)}{\sum_{i=1}^{N}\exp(Sim(Z_{I3D_i}, Z_{Res_j})/\tau)} \quad (2)$$

The total contrastive loss is the sum of two cross-modal and an inter-modal alignments:

$$\mathcal{L}_{\text{pretraining}} = \mathcal{L}_{\text{cross}^{Resnet}} + \mathcal{L}_{\text{cross}^{I3D}} + \mathcal{L}_{\text{inter}} \quad (3)$$

where $\mathcal{L}_{\text{cross}^{Resnet}}$ and $\mathcal{L}_{\text{cross}^{I3D}}$ represent the cross-modal contrastive losses that align the visual representations extracted by the ResNet and I3D branches, respectively, with their corresponding text embeddings. The $\mathcal{L}_{\text{inter}}$ denotes an inter-modal contrastive loss that aims to pull together embeddings from both visual encoders.

3.2 Translation

In this stage, we aim to generate accurate spoken language translations from sign language videos. We employ mBART [18] as the main framework, where we initialize the encoder and decoder with the pretrained shared encoder and text decoder.

Feature Aggregation. After capturing visual features from both visual encoders, a critical step is the fusion of these complementary features. We explore and evaluate various feature aggregation techniques, including element-wise sum, concatenation by channel, and cross-attention as illustrated in Fig. 1.

Visual-Language Adapter. After the aggregation of visual features, they are passed through a visual-language adapter to further refine the aggregated visual features before their integration into the mBART. It contains a fully connected layer, layer normalization, followed by a ReLu activation function.

We use cross-entropy loss to minimize the discrepancies between the predicted and true conditional probabilities of the spoken-language sentence given the sign language video, formally expressed as:

$$\mathcal{L}_{translation} = -logp(S|V) \quad (4)$$

4 Experiments

4.1 Dataset and Evaluation Metrics

Dataset. We evaluate on the Phoenix-2014T [3], which is the most widely used sign language translation dataset. It is a German sign language dataset, containing 7096, 519, and 642 samples for training, dev, and test, with a vocabulary of 2,887 words.

Evaluation Metrics. We use BLEU [21], ROUGE-L [16], and BLEURT [27] metrics to evaluate the performance of our SLT approach. BLEU measures n-gram precision and we use the sacreBLEU [23] implementation. ROUGE-L measures the longest common subsequence between the generated text and the reference, capturing in-order matching words without requiring consecutive overlap. BLEURT [27] is a trained metric that aims to correlate human-quality scoring better. We use the BLEURT-20 checkpoint [24]. Higher scores indicate better translation.

4.2 Implementation Details

Our model is implemented using PyTorch [22] and trained on a single NVIDIA A100 GPU. We resize input sequences into 256×256 which are then cropped into 224×224, where random cropping is used for training and central cropping for inference. Additionally, we apply data augmentation techniques during training with a probability of 50%, including random rotations up to 30°, random resizing with a scale factor of up to 20%, and random translations to shift images up to 10 pixels along both the x and y axes. For pretraining, we employ an SGD optimizer with 0.9 momentum. We initialized the learning rate to 0.02 with a cosine annealing scheduler with batch size 8. For the translation stage, we explore various schedulers detailed in Section 4.4.

We initialized our I3D model with weights pre-trained on the MeineDGS German sign language dataset [15] for isolated sign language translation [31]. For the I3D encoder, we follow the standard setting of processing 16 consecutive frames as input. To extract features over the entire video, we apply a sliding window approach with a stride of 6 frames, resulting in temporal overlaps. To match the temporal resolution with ResNet-based embeddings, we repeat the final feature vector to fill in missing frames at the end of the sequence.

We initialized our mBART model with *mbart-large-cc25*. To save GPU memory, we trim both the model and the tokenizer based on the train set of the dataset.

Table 1: Comparison with the literature on the Phoenix-2014T [3] test set. The best results are bold, and the second-best are underlined.

Method	Publisher	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	BLEURT
Gloss-based							
SL-Transformer [4]	CVPR'20	46.61	33.73	26.19	21.32	-	-
BN-TIN-Transf.+SignBT [40]	CVPR'21	50.80	37.75	29.72	24.32	49.54	-
SLTUNET [38]	ICLR'22	52.92	41.76	33.99	28.47	52.11	-
TwoStream-SLT [7]	NeurIPS'22	54.90	42.43	34.46	28.95	53.48	-
Weakly supervised gloss-free							
VAP [14]	ECCV'24	53.07	-	-	26.16	51.28	-
Gloss-free							
GFSLT-VLP [39]	ICCV'23	43.71	33.18	26.11	21.44	42.49	-
SignCL [37]	NeurIPS'24	**49.76**	**36.85**	**29.97**	22.74	**49.04**	-
Sign2GPT(w/PGP) [34]	ICLR'24	49.54	35.96	28.83	22.52	48.94	-
FLa-LLM [8]	LREC-COLING'24	46.29	35.33	28.03	23.09	45.27	-
SignLLM [11]	CVPR'24	45.21	34.78	28.05	23.40	44.49	-
DVE-SLT (Ours)		<u>49.29</u>	<u>36.68</u>	<u>28.96</u>	**23.81**	<u>48.89</u>	53.59

Table 2: Performance of different visual encoders. BN: BLEU-N, R: Rouge-L, BRT: BLEURT.

Encoder	B1	B2	B3	B4	R	BRT
ResNet [13]	46.68	34.39	26.94	22.11	47.70	53.15
I3D [5]	46.90	34.66	27.41	22.71	46.69	52.96
ResNet + I3D	**49.29**	**36.68**	**28.96**	**23.81**	**48.89**	**53.59**

Table 3: Performance of different feature fusion techniques. BN: BLEU-N, R: Rouge-L, BRT: BLEURT.

Method	B1	B2	B3	B4	R	BRT
Element-wise sum	48.11	35.84	28.17	23.18	48.04	53.20
Channel-wise concat	**49.29**	**36.68**	**28.96**	**23.81**	**48.89**	**53.59**
Cross-attention	48.09	35.89	28.53	23.69	48.74	53.20

4.3 Comparison with State-of-the-art Methods

Table 1 compares our method against existing gloss-based and gloss-free sign language translation approaches on the Phoenix-2014T benchmark. As expected, gloss-based and weakly supervised methods tend to outperform gloss-free methods due to the presence of intermediate supervision or auxiliary data.

In the gloss-free setting, SignCL [37] and Sign2GPT [34] aim to learn effective representations by introducing new contrastive learning or pretraining strategies that achieve better BLEU scores with lower n-gram (BLEU-1 and BLEU-2) and ROUGE scores. On the other hand, recent large language model (LLM) based systems such as FLa-LLM [8] and SignLLM [11], achieve higher BLEU-4 scores, which is the most commonly used translation metric. Notably, our approach outperforms FLa-LLM [8] and SignLLM [11] across all reported metrics, despite relying on a lighter translation model (Ours: mBART [18] with 3 layers vs. mBART with 12 layers or LLaVA [17]). Notably, our model achieves the highest BLEU-4 score (23.81) among all gloss-free methods, outperforming the best competitor, SignCL, by +0.84 BLEU-4. This result highlights the strength of our dual encoder in capturing complementary features, leading to more accurate and fluent sentence-level translations.

4.4 Ablation Studies

Our experiments include assessing the impact of visual encoders, feature aggregation strategies, and optimization techniques on sign language translation performance.

Visual Encoders. To evaluate the contributions of the two visual encoders, we conducted experiments using each encoder individually. As seen in Table 2, the individual performances of ResNet [13] and I3D [5] were found to be comparable, with each encoder showing slight advantages on different metrics. This suggests that both visual encoders effectively capture features relevant to sign language representation. However, fusing them by concatenating in the channel domain enhanced the results.

Feature aggregation techniques. We explore several feature aggregation techniques: element-wise sum, channel-wise concatenation, and cross-attention. As shown in Table 3, concatenation yields the best performance across all metrics. We hypothesize that this is due to its ability to preserve the full information from both modalities. In contrast, element-wise addition may introduce mixing signals or potential information loss. We also experiment with a cross-attention mechanism, which enables the model to attend the most relevant features. However, its additional complexity might hinder optimization, especially when the two visual streams are already semantically aligned via pretraining. Nevertheless, all fusion strategies outperformed using either ResNet or I3D alone, highlighting the benefit of leveraging complementary visual features for translation.

Effect of scheduler. We compare the performance of different learning rate schedulers, including Cosine Annealing (CosAnLR) [19], Exponential (ExpLR), and One Cycle (OneCycleLR) [32], as summarized in Table 4. The results highlight that both CosAnLR and OneCycleLR achieved competitive performance. However, OneCycleLR demonstrates a significant advantage by achieving comparable scores in only 150 epochs, while CosAnLR requires 200 epochs.

Table 4: Performance of different learning rate schedulers and hyperparameters. *CosAnLR, ExpLR, OneCyleLR* stands for Cosine Annealing, Exponential, and One Cycle Learning Rate schedulers. *pct_start* represents the percentage of the training cycle spent increasing the learning rate. *gamma* represents the multiplicative factor of learning rate decay.

Scheduler	# Epoch	Batch Size	Hyperparams	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	BLEURT
CosAnLR	200	4	lr=0.01	47.54	34.99	27.53	22.65	47.77	53.28
CosAnLR	200	8	lr=0.02	48.52	35.91	28.18	23.18	48.24	53.56
ExpLR	200	8	lr=0.02, gamma=0.96	47.76	35.09	27.46	22.61	47.60	51.21
OneCyleLR	150	8	max lr=0.02, pct_start=0.35	**49.29**	**36.68**	**28.96**	**23.81**	**48.89**	**53.39**

This indicates that OneCycleLR can converge faster while maintaining performance. In contrast, the Exponential LR scheduler underperformed in all metrics, suggesting that a fixed decay rate may be suboptimal for our two-stage training pipeline. These findings highlight the importance of selecting an appropriate learning rate scheduler for performance and efficiency.

Qualitative analysis. To better understand how our model attends to relevant signing segments during translation, we visualize attention patterns between video frames and generated tokens. These qualitative results complement our quantitative findings and offer insights into how well the model aligns visual and linguistic representations.

Figure 2 visualizes the cross-attention weights between video frames (y-axis) and generated text tokens (x-axis) during decoding. Brighter regions indicate stronger attention from the decoder to specific frames when predicting the corresponding word. To improve interpretability, we annotate the relevant video frames with the most semantically aligned tokens, using color-coded bounding boxes that match between the video and attention map.

We observe that the decoder assigns high attention scores to the correct signing segments when producing their corresponding words. For example, as shown in Figure 2, the token "montag" (red) aligns with the correct sign for "Monday". The attention is temporally localized and consistent, suggesting that the model learns to ground generated text in the relevant spatio-temporal cues. These results highlight the interpretability of our architecture and its ability to align visual and linguistic semantics without gloss supervision.

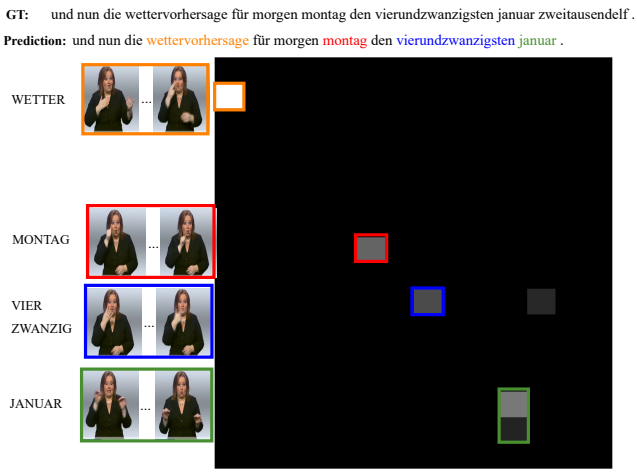

Figure 2: Cross-attention map between video frames (y-axis) and generated tokens (x-axis). Brighter regions indicate higher attention, showing that the model focuses on the correct frames. The matching colors highlight frames and tokens that align with each other.

5 Conclusion

In this paper, we propose a novel dual visual encoder framework, DVE-SLT, for sign language translation. We introduce a dual-objective contrastive alignment strategy during pretraining. This strategy encompasses cross-modal alignment to better correlate visual sign representations with textual descriptions, and inter-modal alignment that encourages consistency between the complementary features extracted by our dual visual encoders. Our approach demonstrates that combining two visual encoders improves the performance of gloss-free sign language translation by enabling the model to capture complementary visual information from different aspects of the sign language video. Future work could explore integrating additional modalities and exploring alternative pretraining strategies to further enhance translation quality.

Ethical Impact Statement

This research aims to enhance sign language translation by leveraging deep learning techniques. While this work has the potential to improve accessibility, we are aware of potential concerns and risks. Since our model is trained on a dataset containing individuals, it is important to address privacy concerns. We used a publicly available dataset, namely Phoenix-2014T [3], which includes consent for non-commercial use. However, we would like to acknowledge that the dataset is limited to weather broadcast content. Although the model demonstrates high performance within this context, it might introduce biases and affect its generalizability to other domains. Additionally, we acknowledge the risk of incorrect translations, which could lead to miscommunication in practical applications. To mitigate this, we focus on obtaining more robust visual features to enhance translation accuracy.

Acknowledgments

We would like to thank Necati Cihan Camgoz for the valuable discussions and feedback. This work was supported by the SNSF project 'SMILE II' (CRSII5 193686), the Innosuisse IICT Flagship (PFFS-21-47), EPSRC grant APP24554 (SignGPT-EP/Z535370/1), 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] 2025. *World Federation of the Deaf.* https://wfdeaf.org/contact/faqs/ Accessed: 2025-05-30.

[2] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, et al. 2021. Bbc-oxford british sign language dataset. *arXiv preprint arXiv:2111.03635* (2021).

[3] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[4] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10023–10033.

[5] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6299–6308.

[6] Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5120–5130.

[7] Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2022. Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems* 35 (2022), 17043–17056.

[8] Zhigang Chen, Benjia Zhou, Jun Li, Jun Wan, Zhen Lei, Ning Jiang, Quan Lu, and Guoqing Zhao. 2024. Factorized Learning Assisted with Large Language Model for Gloss-free Sign Language Translation. *the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)* (2024).

[9] Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. 2023. Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19016–19026.

[10] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15180–15190.

[11] Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. Llms are good sign language translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18362–18372.

[12] Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 297–304.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[14] Peiqi Jiao, Yuecong Min, and Xilin Chen. 2025. Visual Alignment Pre-training for Sign Language Translation. In *European Conference on Computer Vision*. Springer, 349–367.

[15] Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. 2020. MEINE DGS – annotiert. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release / MY DGS – annotated. Public Corpus of German Sign Language, 3rd release. doi:10.25592/dgs.corpus-3.0

[16] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.

[18] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics* 8 (2020), 726–742. doi:10.1162/tacl_a_00343

[19] Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Skq89Scxx

[20] Liliane Momeni, Hannah Bull, KR Prajwal, Samuel Albanie, Gül Varol, and Andrew Zisserman. 2022. Automatic dense annotation of large-vocabulary sign language videos. In *European Conference on Computer Vision*. Springer, 671–690.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[23] Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Association for Computational Linguistics, Belgium, Brussels, 186–191. https://www.aclweb.org/anthology/W18-6319

[24] Amy Pu, Hyung Won Chung, Ankur P Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of EMNLP*.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[26] Ludan Ruan, Anwen Hu, Yuqing Song, Liang Zhang, Sipeng Zheng, and Qin Jin. 2023. Accommodating audio modality in CLIP for multimodal processing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 9641–9649.

[27] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696* (2020).

[28] Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. 2022. Open-Domain Sign Language Translation Learned from Online Video. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 6365–6379. doi:10.18653/v1/2022.emnlp-main.427

[29] Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. 2022. Ttic's wmt-slt 22 sign language translation system. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*. 989–993.

[30] Ozge Mercanoglu Sincan, Necati Cihan Camgoz, and Richard Bowden. 2023. Is context all you need? scaling neural sign language translation to large domains of discourse. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1955–1965.

[31] Ozge Mercanoglu Sincan, Necati Cihan Camgoz, and Richard Bowden. 2024. Using an LLM to Turn Sign Spottings into Spoken Language Sentences. *arXiv preprint arXiv:2403.10434* (2024).

[32] Leslie N Smith and Nicholay Topin. 2019. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, Vol. 11006. SPIE, 369–386.

[33] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[34] Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. Sign2GPT: Leveraging Large Language Models for Gloss-Free Sign Language Translation. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=LqaEEs3UxU

[35] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 6787–6800.

[36] Huijie Yao, Wengang Zhou, Hao Feng, Hezhen Hu, Hao Zhou, and Houqiang Li. 2023. Sign language translation with iterative prototype. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15592–15601.

[37] Jinhui Ye, Xing Wang, Wenxiang Jiao, Junwei Liang, and Hui Xiong. 2024. Improving Gloss-free Sign Language Translation by Reducing Representation Density. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=FtzLbGoHW2

[38] Biao Zhang, Mathias Müller, and Rico Sennrich. 2022. SLTUNET: A Simple Unified Model for Sign Language Translation. In *The Eleventh International Conference on Learning Representations (ICLR)*.

[39] Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 20871–20881.

[40] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving Sign Language Translation With Monolingual Data by Sign Back-Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1316–1325.

[41] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2024. LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=QmZKc7UZCy