Enhancing Indian Sign Language Translation via Motion-Aware Modeling

Anal Roy Chowdhury and Debarshi Kumar Sanyal
Indian Association for the Cultivation of Science, Kolkata, India
analroychowdhury084@gmail.com, debarshi.sanyal@iacs.res.in

Abstract

Sign language translation (SLT) has witnessed rapid progress in the deep learning community across several sign languages, including German, American, British, and Italian. However, Indian Sign Language (ISL) remains relatively underexplored. Motivated by recent efforts to develop large-scale ISL resources, we investigate how existing SLT models perform on ISL data. Specifically, we evaluate three approaches: (i) training a transformer-based model, (ii) leveraging visual-language pretraining, and (iii) tuning a language model with pre-trained visual and motion representations. Unlike existing methods that primarily use raw video frames, we augment the model with optical flow maps to explicitly capture motion primitives, combined with a multi-scale feature extraction method for encoding spatial features (SpaMo-OF). Our approach achieves promising results, obtaining a BLEU-4 score of 8.58 on the iSign test set, establishing a strong baseline for future ISL translation research.

1 Introduction

Sign languages bridge the communication gap between deaf and hearing communities. The World Health Organization[1] predicts that by 2050, over 700 million people will experience disabling hearing loss. This growing prevalence highlights the urgent need for assistive technologies that can support inclusion and accessibility. Sign language translation (SLT) has emerged as a promising research area, with extensive studies on German, American, and British Sign Languages, largely enabled by the availability of large-scale datasets such as PHOENIX-2014T (German) (Camgoz et al., 2018), How2Sign (Duarte et al., 2021) and OpenASL (Shi et al., 2022) (American), and BOBSL (British) (Albanie et al., 2021). SLT methods in the literature typically fall into two categories: gloss-based and gloss-free. Gloss supervision (Camgoz et al., 2020; Chen et al., 2022) has been shown to improve translation quality, but many datasets lack gloss annotations due to their high cost. The shortage of trained sign language experts and the expense of annotation have therefore pushed the community toward gloss-free approaches (Lin et al., 2023; Gong et al., 2024; Wong et al., 2024; Jang et al., 2025; Hwang et al., 2025).

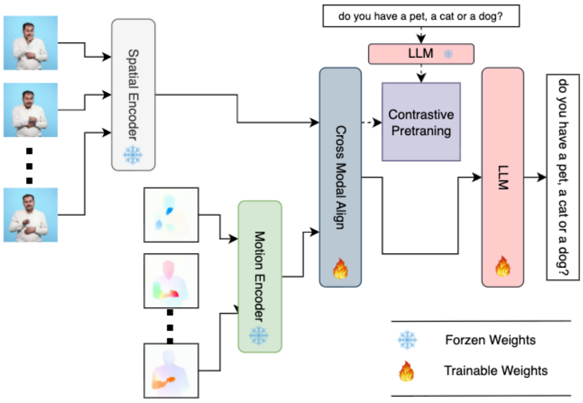

Figure 1: Overall architecture of our SLT framework.

All existing studies have primarily focused on developing translation systems for well-resourced sign languages such as German, Chinese, American, and British. In contrast, low-resource sign languages like Indian Sign Language (ISL) remain largely overlooked. The recent release of large-scale ISL datasets such as ISLTranslate (Joshi et al., 2023) and iSign (Joshi et al., 2024), together with the underperformance of ISL when using existing models, motivates us to examine whether recent architectures that perform well on other sign languages can also generalize to ISL. To this end, we evaluate three representative approaches for SLT: (i) training transformers from scratch (Camgoz et al., 2020), (ii) visual-language pretraining using contrastive learning (Zhou et al., 2023), and (iii) finetuning Large Language Models (LLMs) with multi-scale

[1] https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

Proceedings of the Workshop on Sign Language Processing (WSLP), pages 43–50
December 20-24, 2025 ©2025 Association for Computational Linguistics

spatial features and motion features extracted using a pre-trained backbone. We further adopt the recently proposed SpaMO model (Hwang et al., 2025), which achieves state-of-the-art results on several SLT benchmarks but has not yet been applied to ISL. SpaMO leverages multi-scale feature extraction from input video frames to improve downstream translation performance. We extend it by incorporating optical flow features, which significantly improve results on the iSign dataset by better capturing motion cues in signed gestures. To ensure reliable evaluation, we curated a noise-free subset of iSign and conducted experiments on this data.

Our main contributions are as follows: i) We conduct the first systematic evaluation of three representative SLT approaches on Indian Sign Language. ii) We enhance translation performance by augmenting SpaMO with optical flow features to capture motion primitives alongside multi-scale spatial features. iii) We curate and release a carefully selected subset of iSign to enable robust evaluation for future ISL translation research. The dataset can be accessed using the following link `https://github.com/Analroy/SpaMo-OF.git`.

2 Related Work

The SLT framework has evolved from RNN-based to Transformer-based architectures (Camgoz et al., 2018, 2020), where sequential models take CNN-based features as input. More recent approaches to building better translation systems focus on capturing richer representations, such as pose features or a combination of pose and RGB features (Chen et al., 2022), as well as sign-aware representations (Hu et al., 2021, 2023). To learn stronger sign-specific representations, (Zhou et al., 2023) proposed pretraining the visual encoder, while (Lin et al., 2023) employed contrastive pretraining of the visual encoder using pseudo-gloss supervision. Recent advances in LLMs have also attracted attention (Gong et al., 2024; Wong et al., 2024; Chen et al., 2024), as researchers explore leveraging large-scale pretrained models and adapting them to domain-specific data using parameter-efficient methods such as LoRA (Hu et al., 2022). In contrast to these resource-intensive pretraining approaches, (Hwang et al., 2025) demonstrated that multi-scale features and motion features extracted from a frozen model, when aligned to the LLM space, can achieve improved translation performance by applying LoRA tuning only to the language model.

3 Method

We aim to translate a sign language video $V = [f_1, f_2, \ldots, f_T]$ into a spoken language sentence $Y = [w_1, w_2, \ldots, w_S]$ by leveraging complementary spatial and motion features, aligning them with textual representations, and decoding them with an LLM. The overall pipeline is illustrated in Fig.1.

3.1 Feature Extraction

Spatial Features: A Vision Transformer captures multiscale spatial representations S^2 (Shi et al., 2025) from input frames, following (Hwang et al., 2025). These features encode detailed hand shapes and body postures across scales.

Motion Features: To explicitly capture temporal dynamics, we combine:

Optical Flow: Computed using Global Motion Aggregation (GMA) (Jiang et al., 2021), which robustly handles occluded hand movements (Fig. 2).

VideoMAE Primitives: VideoMAE (Tong et al., 2022) processes 16-frame segments to learn higher-level motion patterns.

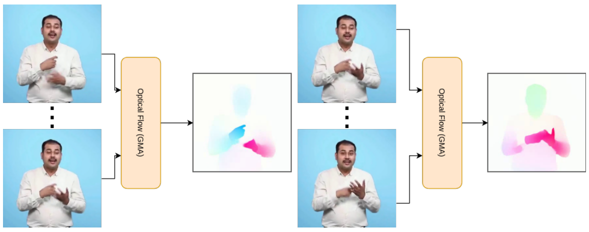

Figure 2: Optical flow map extraction using the GMA.

3.2 Cross-Modal Alignment

Spatial and motion features are projected into the textual embedding space via an alignment module (Hwang et al., 2025), consisting of a linear layer, 1D TCN, and MLP. To bridge the modality gap, we warm up this module with softmax-based contrastive learning (Radford et al., 2021; Jia et al., 2021), aligning embeddings of matching sign-text pairs while pushing apart mismatched ones. Only the alignment module is updated, preserving the LLM's language capabilities and providing well-initialized features for SLT training.

3.3 Language Modeling

Aligned visual and motion features are fed into a multilingual LLM fine-tuned with LoRA to generate the target sentence Y. To focus the LLM on the SLT task, we employ a task-specific prompt (Hwang et al., 2025) that provides a clear instruction, e.g., "Translate the given sentence into Indian," along with multilingual reference translations (e.g., Hindi, French, Spanish) sampled from the training set. The prompt template is shown in Appendix A. Each reference is formatted as $[SRC] = [TRG]$, enabling in-context learning while preventing direct exposure to the target sentence by shuffling pairs during training. At test time, a translation pair from the training set serves as the reference.

4 Experiments

4.1 Datasets

We used the following datasets in our experiments.

German Sign Language (DGS): The RWTH-PHOENIX-2014T (Phoenix-14T) dataset (Camgoz et al., 2018) is the standard benchmark for DGS translation. It contains 7,096 training, 519 validation, and 642 test samples, each aligned with German sentences. The dataset covers weather forecast scenarios interpreted by professional sign language interpreters on television and includes a vocabulary of 2,887 words.

Indian Sign Language (ISL): iSign (Joshi et al., 2024) is a recently introduced large-scale dataset for ISL, comprising over 127k sentence-aligned signing videos collected from diverse real-world contexts.

4.2 Balanced Subset Construction from iSign (ISL)

We retain only sentences containing 5–15 words, reducing the dataset from 127k to 76k samples. Sentences shorter than 4 words are excluded, as very short translations primarily produce low BLEU-1/BLEU-2 scores. When such samples constitute a large portion of the data, the averaged BLEU-4 score no longer reflects meaningful translation quality. Similarly, extremely long sentences, i.e., those with more than 15 words (which have 310 frames on average and up to 2370 frames) are also removed, as their excessive frame counts introduce variability and noise, making model training unstable and inefficient.

From this pool, we construct a balanced subset of 10K samples (avg. 200 frames/sample) as follows: (i) **Word frequency grouping:** vocabulary is split into *rare* (<5 occurrences), *mid-frequency* (5–50), and *common* (>50). (ii) **Sample prioritization:** sentences with rare or mid-frequency words are preferentially selected to ensure coverage of underrepresented words. (iii) **Subset construction:** samples are chosen until the 10K quota is met, filling any gap with random draws. (iv) **Vocabulary coverage:** this guarantees diversity and balance, supporting more effective training.

4.3 Evaluation Metrics

To assess the quality of sign language translations, we employ standard evaluation metrics commonly used in the machine translation literature: BLEU (Papineni et al., 2002) and ROUGE-L (Lin and Och, 2004). BLEU measures n-gram precision by comparing predicted translations with ground-truth references, and we report scores from BLEU-1 through BLEU-4 using the SacreBLEU[2].

4.4 Contending Methods

We evaluate the following SLT models:

SLT (GF) (Camgoz et al., 2020): It is a transformer-based model that jointly learns sign recognition and translation in an end-to-end manner, using CTC loss for alignment. We use the GF framework of this model.

GFSLT-VLP (Zhou et al., 2023): It is a gloss-free framework that combines CLIP-based contrastive learning with masked self-supervised objectives, enabling robust cross-modal representations and strong translation without gloss annotations.

SpaMo (Hwang et al., 2025): It uses off-the-shelf visual encoders for spatial and motion features, combined with language prompts and a lightweight visual-text alignment stage before SLT supervision.

4.5 Implementation Details

We follow the architecture and training setup of SpaMo (Hwang et al., 2025) for spatial–motion feature extraction, cross-modal fusion, and language modeling. Spatial features are obtained from CLIP ViT-L/14 (Radford et al., 2021), while motion representations are enhanced with optical flow maps estimated using global motion averaging (GMA) (Jiang et al., 2021), followed by VideoMAE-L/16 (Tong et al., 2022) over 16-frame clips with a stride of 8. For the language model, we employ Flan-T5-XL (Chung et al., 2024) with LoRA adaptation, using a 1K-step warm-up on both Phoenix-14T and iSign. All experiments are conducted on a single NVIDIA A100 GPU.

[2] https://github.com/mjpost/sacrebleu

4.6 Results

We present our experimental results in Table 1, comparing three existing models and our approach on the Phoenix-14T and iSign-10k datasets.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGH-L
Phoenix-14T					
SLT (GF)	44.99	32.14	24.62	20.00	**45.32**
GFSLT-VLP	39.52	29.15	22.54	18.23	38.60
SpaMo	**46.41**	**33.15**	**25.22**	**20.18**	42.21
SpaMo-OF (ours)	38.06	25.26	18.58	14.72	33.52
iSign-10k					
SLT (GF)	9.84	2.87	1.33	0.73	9.08
GFSLT-VLP	9.24	4.10	2.56	1.93	9.24
SpaMo	25.42	12.90	9.13	7.35	16.23
SpaMo-OF (ours)	**27.91**	**15.00**	**10.67**	**8.58**	**18.98**

Table 1: Performance of SLT methods on the **Phoenix-14T** and **iSign-10k test sets**, reported in BLEU and ROUGE-L.

Results on Phoenix-14T: We used the preprocessed data from (Camgoz et al., 2020) to conduct experiments with SLT(GF). For the other models, we had to prepare the dataset in the appropriate format. The findings, shown in Table 1, indicate that (Hwang et al., 2025) achieves the best performance, with a BLEU-4 score of 20.18. Incorporating optical flow maps (SpaMo-OF) results in a performance drop.

Results on iSign-Full: We conducted experiments only with SLT (GF) (Camgoz et al., 2020) for this dataset. We preprocessed the dataset in a manner consistent with Phoenix-14T. Features were extracted from the signing videos using EfficientNet (Tan and Le, 2019). But this yielded poor performance (BLEU-4: 0.32), similar to the trend reported by (Joshi et al., 2024). Given that the dataset contains over 127k samples, further experimentation with other models proved impractical within our resources due to the substantial computational resources required.

Results on iSign-10k: Our experiments on the iSign-10k subset suggest that incorporating optical flow maps enables the model to leverage occluded motion cues more effectively, as captured by GMA (Jiang et al., 2021). As shown in Table 1, our method achieves a BLEU-4 score of 8.58, the highest among all evaluated approaches.

As part of our ablation study, we examined the effect of in-context examples on both datasets and observed marginal performance gains, with three in-context examples yielding the best BLEU-4 scores. We further evaluated the impact of in-context examples during test time. Detailed results are provided in Appendix B.

4.7 Qualitative Evaluation of Translation Results

The translation system achieves mixed performance with near-perfect accuracy on simple sentences with common vocabulary but struggles with sentences containing numbers and technical references. A few examples of correct and incorrect translations produced by SpaMo-OF on iSign-10k test set are shown in Table 2. More qualitative examples and detailed error cases are provided in Appendix C.

Ground Truth: Generated:	when he began to sing, the air became warm. when he began to sing, the air became warm.
Ground Truth: Generated:	once a farmer and his wife lived in a village with their small son. once a farmer and his wife lived in a village with their small son.
Ground Truth:	soldiers were paid regular salaries and maintained by the king throughout the year.
Generated:	soldiers were paid regular salaries and maintained by the king throughout the year.
Ground Truth: Generated:	do you have a pet, a cat or a dog? do you have a pet, a cat or a dog?
Ground Truth: Generated:	she was also a certified flight instructor. after qualifying as a pilot, kalpana was born in karnal, haryana.
Ground Truth: Generated:	story, mittu and the yellow mango. the peacock is blue and green.
Ground Truth: Generated:	many people feel bewildered by the speed of technological innovation. the company is aiming to become a global player in the industry.
Ground Truth: Generated:	look at figure 7.25a and b carefully. identify the parts of the pistol with the help of figure 7.24.

Table 2: Translation examples from the iSign-10k test set using SpaMo-OF. Blue indicates partial matches; top rows show correct outputs, while bottom rows illustrate common errors.

5 Conclusion

This work presented the first systematic evaluation of representative SLT models on Indian Sign Language, highlighting the challenges of extending methods successful in well-resourced languages to a low-resource setting. We have shown that curating a clean, balanced subset of iSign is critical for reliable evaluation and that augmenting SpaMo with optical flow features yields notable improvements, achieving a BLEU-4 score of 8.58. Our results suggest that dataset quality, rather than scale alone, is key to translation performance, and that motion-aware representations play an essential role in modeling signed communication. Future efforts should focus on constructing cleaner benchmarks and designing models that more effectively integrate spatial and motion primitives to advance robust ISL translation.

Limitations

Our work is limited by computational resources, which prevented training on the full iSign dataset (127k+ samples). We relied on a curated 10k subset for feasibility. Optical flow computation also adds overhead, limiting scalability. Additionally, the curated subset may not capture the full linguistic diversity of ISL. Future work should explore more efficient architectures and larger, more diverse datasets to improve performance and generalization.

References

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. 2021. BOBSL: BBC-Oxford British Sign Language Dataset. *arXiv*.

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)*.

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 (CVPR)*.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie LIU, and Brian Mak. 2022. Two-Stream Network for Sign Language Recognition and Translation. In *Advances in Neural Information Processing Systems*, volume 35, pages 17043–17056. Curran Associates, Inc.

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. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7071–7081. ELRA and ICCL.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

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)*, pages 18362–18372.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. 2023. SignBERT+: Hand-Model-Aware Self-Supervised Pre-Training for Sign Language Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11221–11239.

Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. 2021. SignBERT: Pre-Training of Hand-Model-Aware Representation for Sign Language Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11087–11096.

Eui Jun Hwang, Sukmin Cho, Junmyeong Lee, and Jong C. Park. 2025. An Efficient Gloss-Free Sign Language Translation Using Spatial Configurations and Motion Dynamics with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3901–3920. Association for Computational Linguistics.

Youngjoon Jang, Haran Raajesh, Liliane Momeni, Gül Varol, and Andrew Zisserman. 2025. Lost in Translation, Found in Context: Sign Language Translation with Contextual Cues. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 8742–8752.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. 2021. Learning to estimate hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9772–9781.

Abhinav Joshi, Susmit Agrawal, and Ashutosh Modi. 2023. ISLTranslate: Dataset for translating Indian Sign Language. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages

10466–10475. Association for Computational Linguistics.

Abhinav Joshi, Romit Mohanty, Mounika Kanakanti, Andesha Mangla, Sudeep Choudhary, Monali Barbate, and Ashutosh Modi. 2024. iSign: A Benchmark for Indian Sign Language Processing. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10827–10844. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. Gloss-Free End-to-End Sign Language Translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12904–12916. Association for Computational Linguistics.

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*, pages 311–318.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. 2025. When do we not need larger vision models? In *Computer Vision – ECCV 2024*, pages 444–462. Springer Nature Switzerland.

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*, pages 6365–6379, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pretraining. In *Advances in Neural Information Processing Systems*, volume 35, pages 10078–10093. Curran Associates, Inc.

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*.

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)*, pages 20871–20881.

Appendix

A Prompt Template

In this section, we describe the prompt template used in our experiments for sign language translation. To facilitate multilingual in-context learning, we leverage the Google Translate API[3] to translate Indian Sign Language (ISL) sentences into multiple target languages (Hindi, Spanish, and French), enabling the model to benefit from cross-lingual cues.

Sign Video Input:	[Extracted Sign Feature]
Instruction:	Translate the given sentence into Indian.
In-context Examples:	पेड़ का रंग क्या है?
	¿Cuál es el color del árbol?
	Quelle est la couleur de l'arbre?

Table 3: Example of the prompt format used in our experiment.

B Ablation Study

Table 4 shows the impact of varying the number of in-context examples during training on the Phoenix-14T and iSign-10k datasets. We observe that increasing the number of examples leads to consistent, albeit modest, gains across all BLEU and ROUGE metrics. Interestingly, the best performance—reflected in the highest BLEU-4 and ROUGE-L scores—is achieved with three in-context examples, indicating that a small amount of contextual guidance can effectively enhance the model's ability to align signs with corresponding text.

[3] https://cloud.google.com/translate?hl=en

No. of in-context examples	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
Phoenix-14T					
0	35.14	22.03	15.65	12.20	29.38
1	33.52	20.68	14.41	11.02	28.93
2	35.79	22.99	16.51	12.91	31.31
3	**37.45**	**24.37**	**18.01**	**14.41**	**32.63**
iSign-10k					
0	26.61	14.67	10.67	8.73	17.75
1	26.22	13.79	9.89	8.08	16.93
2	26.26	13.98	10.06	8.23	16.94
3	**27.95**	**15.37**	**11.03**	**8.92**	**18.99**

Table 4: Performance with varying numbers of in-context examples during training (all models tested with zero in-context examples).

Table 5 examines the effect of in-context examples during testing, with all models trained using three examples. For Phoenix-14T (German), the in-context examples are in English, Spanish, and French, while for iSign-10k (ISL), they are in Hindi, Spanish, and French. The results indicate that providing a small number of in-context examples can slightly improve performance. Notably, for iSign-10k, including Hindi examples at test time appears to enhance translation quality, suggesting that using a language closely related to the target output can help the model better generalize, whereas adding more examples beyond two does not consistently yield further gains.

No. of in-context examples	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
Phoenix-14T					
0	35.70	22.87	16.63	13.15	31.34
1	**37.63**	**24.60**	**18.10**	14.37	**32.82**
2	37.30	24.22	17.77	14.18	32.37
3	37.45	24.37	18.01	**14.41**	32.63
iSign-10k					
0	27.91	15.00	10.67	8.58	18.98
1	28.11	15.51	11.22	**9.12**	**19.07**
2	**28.14**	**15.62**	**11.25**	9.09	18.93
3	27.95	15.37	11.03	8.92	18.99

Table 5: Performance with varying numbers of in-context examples during testing (all models trained with 3 in-context examples).

C Qualitative Evaluation of Translation Results

Tables 6, 7, and 8 provide additional translation examples drawn from iSign-10k test set, with generated sentences produced using SpaMo-OF. The tables contain both high-quality and poor-quality system outputs. Blue highlighting indicates exact word matches between ground truth and generated text, revealing the system's binary performance pattern with either extensive lexical preservation or complete semantic divergence.

Ground Truth:	why did the farmer's wife strike the mongoose with her basket?
Generated:	why did the farmer's wife strike the mongos with her basket?
Ground Truth:	saida's mother had been ailing for a long time. fever, cough.
Generated:	saida's mother had been ailing for a long time. fever, cough.
Ground Truth:	what did raja nanda do to anger the court of gautamiputra?
Generated:	prashastis and what they tell us?
Ground Truth:	in the city of madurai, there was a craftsperson named chandragupta.
Generated:	between 2,200 and 1900 years ago. between 2,200 and 1900 years ago.
Ground Truth:	taxes on crops were the most important. this was because most people were farmers.
Generated:	usually the tax was fixed at 1/6th of what was produced.
Ground Truth:	connect the dots to write a circle.
Generated:	say which one will float and which one will sink.
Ground Truth:	the mauryan rulers did not try to control these changes.
Generated:	9 years after becoming king, I conquered.
Ground Truth:	when people began writing on cloth.
Generated:	trace the river indus and its tributaries in the map.
Ground Truth:	including craftspersons and merchants. in addition,
Generated:	for instance, inscriptions from matura mention goldsmiths.
Ground Truth:	a poor shepherd once lived in iran. though uneducated,
Generated:	a poor shepherd once lived in iran. he became famous for his piousness.

Table 6: Additional translation examples showing mixed quality patterns including near-matches, entity disambiguation failures, instruction drift, and complete semantic divergence. Blue highlights exact word matches between ground truth and generated text.

Ground Truth:	he was ordered to explain why he always carried an iron chest.
Generated:	he was ordered to explain why he always carried an iron chest.
Ground Truth:	but an old blanket was all that came out.
Generated:	but an old blanket was all that came out.
Ground Truth:	now the synopsis of the next part of the story.
Generated:	now the synopsis of the next part of the story.
Ground Truth:	which had to flee from its post. the rays got through.
Generated:	which had to flee from its post. the rays got through.
Ground Truth:	day by day, i float my paper boats.
Generated:	day by day, i float my paper boats.
Ground Truth:	we apologize for any inconvenience caused during the repairs.
Generated:	we apologize for any inconvenience caused by the late arrival of train.
Ground Truth:	the little group of travellers scattered in fear but swami haridas called them together.
Generated:	the little group of travellers scattered in fear but swami haridas called them together.
Ground Truth:	there once lived a bird and two newborn babies blue in a forest.
Generated:	there once lived a bird and two newborn babies in a forest.
Ground Truth:	question number five. did she repent her hasty action?
Generated:	question number five. did she repent her hasty action?
Ground Truth:	frankly, i don't like to leave the child alone with the mongoose.
Generated:	i don't like to leave the child alone with the mongoose.

Table 7: Translation examples demonstrating high-quality outputs with perfect or near-perfect semantic preservation, minor lexical variations, and contextual consistency. Blue indicates exact word-level matches between ground truth and generated text.

Content Type	**Example**	**Ground Truth**	**Generated**
		High Performance: Narrative & Simple Sentences	
Narrative	Story introduction	once a farmer and his wife lived in a village with their small son.	once a farmer and his wife lived in a village with their small son.
Narrative	Story continuity	there once lived a bird and two newborn babies in a forest.	there once lived a bird and two newborn babies in a forest.
Historical	Factual statement	soldiers were paid regular salaries and maintained by the king throughout the year.	soldiers were paid regular salaries and maintained by the king throughout the year.
Question	Direct question	do you have a pet, a cat or a dog?	do you have a pet, a cat or a dog?
		Low Performance: Educational & Instructional Content	
Instruction	Drawing activity	connect the dots to write a circle.	say which one will float and which one will sink.
Technical Ref.	Figure reference	look at figure 7.25a and b carefully.	identify the parts of the pistol with the help of figure 7.24.
Historical Ed.	Context instruction	when people began writing on cloth.	trace the river indus and its tributaries in the map.
Entity Ref.	Historical query	what did raja nanda do to anger the court of gautamiputra?	prashastis and what they tell us?
Educational	Historical context	in the city of madurai, there was a craftsperson named chandragupta.	between 2,200 and 1900 years ago. between 2,200 and 1900 years ago.
Biographical	Career context	she was also a certified flight instructor. after qualifying as a pilot,	kalpana was born in karnal, haryana.

Table 8: Translation performance across content types in iSign-10k test set using SpaMo-OF. Blue indicates exact matches. Top section shows perfect translations on narrative and simple content while bottom section reveals failures on relatively more complex content.