ReSLT: Retrieval-enhanced Sign Language Translation with LLMs

Anonymous ACL submission

Abstract

Gloss-free Sign Language Translation (SLT) aims to directly translate visual expressions into spoken language, bypassing intermediate gloss annotations. Recent studies have demonstrated remarkable performance by leveraging Large Language Models (LLMs) in gloss-free SLT. However, existing approaches often fail to fully exploit the potential of LLMs due to simplistic prompt design. To address this gap, we propose **ReSLT**, a **Retrieval-Augmented Generation SLT framework** that utilizes pre-existing linguistic knowledge to enable LLMs to effectively comprehend sign languages. ReSLT incorporates a semantic prompting strategy, aligning video and text embeddings to construct context-aware prompts. Additionally, the proposed framework maintains a lightweight structure, allowing for easy integration with other SLT models, thus enhancing the applicability of LLMs in SLT. Our experiments demonstrate that even with the simplest architecture, ReSLT achieves performance gains in Korean Sign Language and German Sign Language, highlighting its effectiveness and scalability.

1 Introduction

Sign language is a rich and structured visual language that is essential to Deaf communities. However, it remains underexplored in natural language processing (NLP) (Kim et al., 2024a). Its inherently multimodal nature—spanning hand gestures, facial expressions, and body posture—poses unique challenges, as it lacks direct syntactic alignment with spoken or written language. Gloss[1] annotations offer a useful linguistic abstraction, but they are labor-intensive and difficult to scale (Yin and Read, 2020). Consequently, recent research(Zhou et al., 2023; Chen et al., 2024; Wong et al., 2024; Gong et al., 2024; Hwang et al., 2024; Kim et al., 2024b) has shifted toward direct Sign-to-Text translation approaches.

Large Language Models (LLMs)(Chowdhery et al., 2023; Chung et al., 2024; Grattafiori et al., 2024; Yang et al., 2025), pretrained on multilingual corpora, show promise in low-resource translation (Yang et al., 2023). Their ability to model cross-linguistic structures allows generalization with minimal supervision (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023). Due to limited datasets and sparse domain coverage, sign languages are considered low-resource. This has motivated recent efforts to apply LLMs to SLT via few-shot prompting—embedding a small number of translation examples within the prompt. However, existing methods often ignore semantic similarity when selecting examples, which may hinder LLM performance (Rubin et al., 2021). In SLT, where subtle visual variations carry semantic weight, irrelevant prompts can act as noise.

We introduce **ReSLT**, **a retrieval-augmented generation framework** for gloss-free SLT that injects semantically aligned multilingual examples into prompts. For a given sign video, ReSLT retrieves **semantically similar spoken-language sentences** and uses them as **in-context translation examples.** This guides decoding by grounding unfamiliar inputs in familiar linguistic structures. ReSLT is lightweight, adding only a retrieval module to standard LLM-based SLT systems. Despite its simplicity, it surpasses strong baselines on German and Korean SLT and generalizes across domains. Our results show that semantically informed prompting improves LLMs' ability to handle low-resource sign languages.

2 Related Work

2.1 Core Components for Gloss-Free SLT

Gloss-free SLT systems typically consist of (1) a visual feature extractor, (2) a modality adapter,

[1] Gloss represents sign language in writing, connecting signs to their meanings.

and (3) a language model. Feature extractors such as (2+1)D CNNs are widely used for balancing efficiency and temporal modeling (Zhou et al., 2023; Cui et al., 2019). The modality adapter (e.g., MLP or Q-former (Zhang et al., 2024)) projects visual features into the language model's embedding space. We follow this standard pipeline, integrating a semantic retriever to isolate the effect of our prompting strategy.

2.2 Representation Learning in SLT

Aligning visual and linguistic modalities is central in SLT. Prior works(Zhou et al., 2023; Gan et al., 2023; Ye et al., 2024; Hwang et al., 2024; Kim et al., 2024b) uses contrastive learning to embed videos and texts into a shared space. This not only aids translation but also enables semantic retrieval. We adopt this setup to support semantically guided prompting without altering the SLT training objective.

2.3 Prompt Strategies for LLM-Based SLT

Recent SLT work incorporates LLMs via few-shot multilingual prompts, often selected at random (Hwang et al., 2024; Gong et al., 2024). Yet, LLMs are sensitive to the content and order of in-context examples (Lewis et al., 2020; Liu et al., 2021; Batheja and Bhattacharyya, 2023; Winata et al., 2023; Baumann et al., 2024), and poorly chosen prompts can degrade performance (Gao et al., 2020). This underscores the need for semantically grounded prompting—especially for sign languages, which remain largely unfamiliar to most LLMs.

3 Method

We propose ReSLT, a retrieval-augmented generation (RAG) framework that enables LLMs to effectively interpret low-resource sign languages by leveraging pretrained linguistic knowledge. The overall framework is shown in Figure 1. Given a sign video $V = (I_1, I_2, \dots, I_N)$ of N frames, the goal of gloss-free SLT is to generate a spoken-language sentence $S = (W_1, W_2, \dots, W_U)$ of U tokens. ReSLT builds on a minimal framework with a Sign Embedder and a pretrained LLM, adding a Video-to-Text Retriever to examine the effect of semantic prompting. The framework can be easily integrated into existing LLM-based SLT systems.

3.1 Sign Embedder

To effectively interface sign language input with a pretrained LLM, we first encode the visual signal into a compact, temporally-aware representation. We employ a frozen visual backbone (e.g., He et al., 2016; Radford et al., 2021) to encode each frame I_i into visual features $f_i \in \mathbb{R}^D$, which are stacked to form a sequence $F = (f_1, f_2, \dots, f_N)$. We then apply a 1D-CNN to capture short-range temporal dependencies and reduce the sequence length by a factor of 4. The resulting feature sequence is projected via an MLP into the LLM embedding space, yielding sign tokens $F_s = (f_{s1}, f_{s2}, \dots, f_{sN/4}) \in \mathbb{R}^{D'}$.

3.2 Video-Text Aligment

To enable the retrieval of semantically relevant pairs across modalities, we align video and text embeddings in a shared semantic space using a symmetric contrastive loss. Given a mini-batch of video-text pairs $\{(v_j, t_j)\}_{j=1}^{|B|}$, we derive the sign embedding $v_j = \text{AvgPool}(F_{s\{j\}})$ and text embedding $t_j = \text{AvgPool}(E_w(\text{Tokenizer}(Y_j)))$, where Y_j is the target translation text and E_w is the pretrained LLM's embedding layer. The loss is:

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{2|B|} \sum_{j=1}^{|B|} \Bigg[-\log \frac{\exp(\text{sim}(v_j, t_j)/\tau)}{\sum_{k=1}^{|B|} \exp(\text{sim}(v_j, t_k)/\tau)} - \log \frac{\exp(\text{sim}(t_j, v_j)/\tau)}{\sum_{k=1}^{|B|} \exp(\text{sim}(t_j, v_k)/\tau)} \Bigg] \quad (1)$$

where $\text{sim}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$ and τ is a temperature parameter. This training encourages semantically matched video-text pairs to lie close in a shared semantic space, enabling cross-modal retrieval for prompt construction.

3.3 Video-To-Text Retrieval

During both SLT training and inference, the averaged sign embedding v is used to retrieve semantically similar sentences from a multilingual vector database built from the training set. Each entry consists of a key(target-language sentence embedding)-metadata(target text translations in multiple languages), grouped to align with the LLM's prior distribution.

This multilingual knowledge helps the LLM ground unfamiliar sign language inputs by anchoring them to semantically related linguistic expressions in familiar patterns. All text embeddings are computed using the LLM's token embedding layer E_w with average pooling, and cosine similarity is

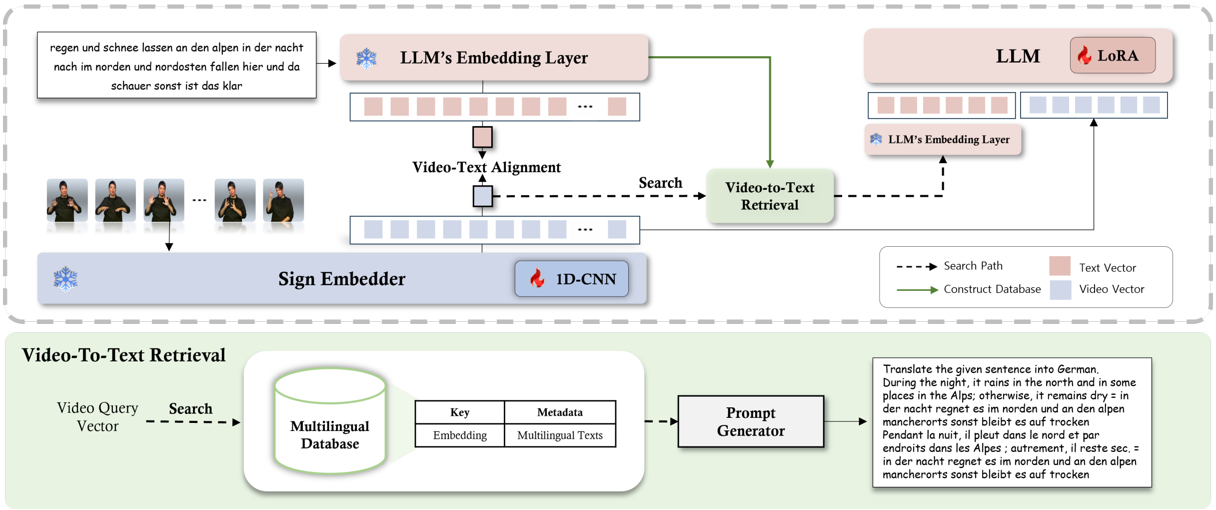

Figure 1: An overview of the ReSLT framework, which consists of three parts: (1) **Sign Embedder** transforms sign video into LLM-compatible token embeddings using a visual encoder and temporal projection. (2) **Video-To-Text Retrieval** retrieves semantically similar multilingual examples using sign embeddings, and constructs prompts via a prompt generator to guide LLM translation, as illustrated in the bottom figure. (3) **LLM** generates translations from sign tokens using prompts and is fine-tuned with LoRA to adapt to the sign language domain.

used for retrieval. To prevent label leakage, ground-truth sentences are excluded from retrieval during training. At inference time, retrieval is restricted to the training set to reflect realistic deployment conditions.

A prompt generator formats the top-k retrieved entries into a prompt containing a translation instruction and multilingual few-shot examples. This prompt P, combined with the sign tokens F_s, guides LLM decoding.

3.4 Large Language Model

To leverage pretrained language knowledge while enabling domain-specific adaptation, we apply LoRA (Low-Rank Adaptation) (Hu et al., 2022) to the LLM. During decoding, the model receives the constructed prompt P followed by the sign tokens F_s. The objective is to minimize the cross-entropy loss between the generated sequence $\hat{y}$ and reference translation y:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{T}\sum_{t=1}^{T} \log P(y_t \mid y_{<t},\ P, F_s) \quad (2)$$

Our framework enables sign language translation by incorporating semantically relevant multilingual examples, requiring only the addition of a retrieval module to existing LLM-based translation frameworks. See Appendix A for implementation details.

4 Experiment

Datasets. We evaluate our method on both Korean and German Sign Language datasets. For **Korean Sign Language (KSL)**, we use dataset provided by the National Institute of Korean Language[2], applied the preprocessing method proposed in the SSL(Kim et al., 2024c). For **German Sign Language (DGS)**, we utilize the RWTH-PHOENIX-Weather 2014T(Camgoz et al., 2018). A detailed description is provided in the Appendix B.

Evaluation Metrics. We use BLEU(Papineni et al., 2002), ROUGE-L(Lin, 2004), and BLEURT(Sellam et al., 2020), widely used in SLT

4.1 Effects of semantic prompting

Lang	type	B1 ↑	B2 ↑	B3↑	B4 ↑	R ↑	BLT ↑
De	Zero	45.62	34.89	27.57	22.71	**45.18**	0.55
	Rand	44.15	33.79	27.12	22.60	43.44	0.55
	Sim	**46.08**	**35.30**	**28.07**	**23.20**	44.73	**0.57**
Ko	Zero	38.77	26.05	18.21	13.16	36.89	0.67
	Rand	38.20	25.80	18.11	13.10	36.35	0.67
	Sim	**38.98**	**26.24**	**18.44**	**13.35**	**37.06**	0.67

Table 1: Evaluation results on the DGS and KSL Sign Language datasets using three prompting strategies: Zero (no examples), Rand (random multilingual examples), and Sim (retrieval-based examples, ours). Metrics include BLEU-1 to BLEU-4, ROUGE-L, and BLEURT.

We evaluate the impact of semantic prompting by comparing three setups: **Zero**, **Rand**, and **Sim(Ours)**. Results across both DGS and KSL are shown in Table 1. Our method consistently outperforms the baselines, achieving up to +0.49 BLEU-4

[2]https://www.korean.go.kr/

and +0.02 BLEURT over Zero in DGS, and showing stable gains in KSL. Notably, **Rand** underperforms **Zero**, indicating that irrelevant prompts degrade performance. These results highlight that semantic relevance in few-shot prompts is crucial for enhancing translation quality—especially in low-resource, non-textual modalities such as sign languages. Qualitative results are in Appendix C.

4.2 Comparison with State-of-the-Art

Lang	Methods	Vis Mod.	LM Size	B1	B2	B3	B4	R
DE	GFSLT(Zhou et al., 2023)	Y	610M	43.71	33.18	26.11	21.44	42.49
	FLa-LLM(Chen et al., 2024)	Y	610M	46.29	35.33	28.03	23.09	45.27
	Sign2Gpt(Wong et al., 2024)	Y	1.7B	49.54	35.96	28.83	22.52	48.90
	SignLLM(Gong et al., 2024)	Y	7B	45.21	34.78	28.05	23.40	44.49
	SpaMo(Hwang et al., 2024)	Y	3B	49.80	37.32	29.50	24.32	46.57
	MMSLT(Kim et al., 2024b)	Y	8B	48.92	38.12	30.79	25.73	47.97
	ours	N	3B	**46.08**	**35.3**	**28.07**	**23.2**	**44.73**
KO	*SLRT(Camgoz et al., 2020)	N	580M	27.39	17.17	11.20	7.57	27.71
	*GFSLT(Zhou et al., 2023)	Y	610M	25.77	15.77	10.03	7.85	26.52
	ours	N	3B	**38.98**	**26.24**	**18.44**	**13.35**	**37.06**

Table 2: Comparison of methods on the DGS and KSL datasets in terms of model size, visual modification, and evaluation metrics. Asterisks (*) denote reproduced results. Our results are highlighted as **bold**, and the best results are underlined.

Table 2 compares our approach to recent SLT systems. Existing work often scales LLMs to larger sizes or modifies the visual encoder with task-specific pretraining and architectural changes. In contrast, we adopt lightweight yet flexible framework - a frozen vision backbone, a retrieval module, and LoRA-based adaptation of a moderately sized LLM.

Since only two model(*) provide released code, we reproduce baseline setups to the best of our ability for KSL. Despite its simplicity, our method achieves competitive performance across both DGS and KSL. Notably, we exceed reproduced baselines on KSL, which spans diverse domains. These results show that competitive SLT performance can be achieved with simple integration of a semantic prompt.

4.3 Impact of Retriever Performance

To isolate the effect of retrieval quality at inference time, we fix the training setup with consistently high-quality examples and vary only the retriever checkpoint during inference (Figure 2). As retrieval accuracy improves in DGS, BLEU-4 scores correspondingly. Although the gains are modest, they are solely attributable to improved retrieval at inference—highlighting the decoder's sensitivity to semantic prompting. Importantly, this decoupling between training and inference enables post hoc retriever upgrades—facilitating lightweight, scalable

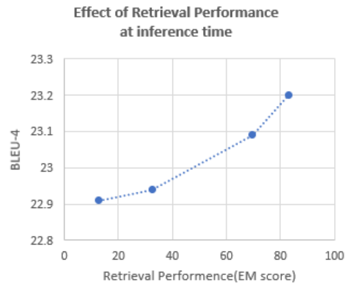

Figure 2: Impact of retrieval quality at inference time on BLEU-4 scores in DGS.

enhancement without end-to-end retraining.

4.4 Cross-Domain Performance Comparison

Type	Tourism	Public Services	Shopping	Healthcare
Zero	15.68	**12.54**	13.15	7.41
Random	15.76	12.08	13.40	8.12
Sim	**15.77**	12.53	**13.68**	**11.18**

Table 3: BLEU-4 scores across four KSL subdomains—Tourism, Public Services, Shopping, and Healthcare—indicate that our method yields substantial improvements in the specialized domain of Healthcare.

We evaluate domain generalization by measuring BLEU-4 across four KSL subdomains: Tourism, Public Services, Shopping, and Healthcare (Table 3). In general-purpose domains, the average performance difference among the three prompting strategies is relatively small, about 0.22. However, in the Healthcare domain, which is characterized by a high density of specialized terminology (e.g., "glycated hemoglobin," "thyroid hormones"), Sim method achieves a notable gain +3.06. These results indicate that semantically grounded prompting becomes valuable as domain complexity and terminology density rise, reinforcing the importance of semantic retrieval in specialized domain.

5 Conclusion

In this work, we introduced ReSLT, designed to address the challenges of gloss-free SLT. Unlike prior approaches that have not placed significant emphasis on prompt design, We leverages semantic retrieval to construct prompts with semantically aligned multilingual examples. This strategy yields competitive results with the simple integration of retrieval for constructing semantic prompts within a minimalistic framework. We explored how LLMs can be effectively utilized in SLT, opening a new direction for maximizing their contextual capabilities.

6 Limitations

While ReSLT demonstrates its effectiveness in gloss-free SLT by achieving notable performance gains, certain limitations remain. First, our evaluation is limited to a single model per language, primarily due to computational constraints and access to extensive pretraining corpora. This choice is not intended to imply that ReSLT is narrowly tailored to specific LLMs, but rather to establish a baseline framework that can be extended to broader model configurations and language scales in future work. Further exploration of multiple LLM architectures with diverse training data would provide a more comprehensive understanding of ReSLT's robustness and generalizability in SLT tasks. Additionally, incorporating models with different parameter scales could reveal how retrieval-based prompting interacts with model capacity, further elucidating the scalability of our approach.

Furthermore, we employ a fixed structure for multilingual prompts, where the number and order of language components are predefined based on rule-based configurations. Despite achieving strong results with this structure, it may not fully capture optimal language combinations or prompt structures for varying SLT contexts. The rigidity of the setup could potentially limit the framework's adaptability to more specialized or emerging sign languages, where linguistic patterns may differ significantly from mainstream datasets. Investigating more adaptive prompting strategies—considering factors such as linguistic similarity, domain specificity, and the inclusion of diverse examples—could further refine retrieval and translation accuracy without compromising the fundamental simplicity of the proposed framework.

References

Akshay Batheja and Pushpak Bhattacharyya. 2023. "a little is enough": Few-shot quality estimation based corpus filtering improves machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14175–14185, Toronto, Canada. Association for Computational Linguistics.

Nils Baumann, Juan Sebastian Diaz, Judith Michael, Lukas Netz, Haron Nqiri, Jan Reimer, and Bernhard Rumpe. 2024. Combining retrieval-augmented generation and few-shot learning for model synthesis of uncommon dsls. In *Modellierung 2024 Satellite Events*, pages 10–18420. Gesellschaft für Informatik eV.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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*, pages 7784–7793.

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)*.

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. *arXiv preprint arXiv:2403.12556*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Runpeng Cui, Hu Liu, and Changshui Zhang. 2019. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, 21(7):1880–1891.

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Kang Xia, Lei Xie, and Sanglu Lu. 2023. Contrastive learning for sign language recognition and translation. In *IJCAI*, pages 763–772.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

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*, pages 18362–18372.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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*, pages 770–778.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Eui Jun Hwang, Sukmin Cho, Junmyeong Lee, and Jong C Park. 2024. An efficient sign language translation using spatial configuration and motion dynamics with llms. *arXiv preprint arXiv:2408.10593*.

Jung-Ho Kim, Changyong Ko, Mathew Huerta-Enochian, and Seung Yong Ko. 2024a. Shedding light on the underexplored: Tackling the minor sign language research topics. In *SIGNLANG*.

Jungeun Kim, Hyeongwoo Jeon, Jongseong Bae, and Ha Young Kim. 2024b. Leveraging the power of mllms for gloss-free sign language translation. *arXiv preprint arXiv:2411.16789*.

Wooyoung Kim, TaeYong Kim, Byeongjin Kim, Myeong Jin MJ Lee, Gitaek Lee, Kirok Kim, Jisoo Cha, and Wooju Kim. 2024c. Korean disaster safety information sign language translation benchmark dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9948–9953.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Genta Indra Winata, Liang-Kang Huang, Soumya Vadlamannati, and Yash Chandarana. 2023. Multilingual few-shot learning via language model retrieval. *arXiv preprint arXiv:2306.10964*.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. Sign2gpt: Leveraging large language models for gloss-free sign language translation. *arXiv preprint arXiv:2405.04164*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages. *arXiv preprint arXiv:2305.18098*.

Jinhui Ye, Xing Wang, Wenxiang Jiao, Junwei Liang, and Hui Xiong. 2024. Improving gloss-free sign language translation by reducing representation density. *arXiv preprint arXiv:2405.14312*.

Kayo Yin and Jesse Read. 2020. Better sign language translation with stmc-transformer. In *International Conference on Computational Linguistics*.

Qiming Zhang, Jing Zhang, Yufei Xu, and Dacheng Tao. 2024. Vision transformer with quadrangle attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3608–3624.

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*, pages 20871–20881.

A Implementation Details

A.1 Framework Detail

Stage 1 Visual features were extracted from individual frames of the sign language videos using the pretrained CLIP ViT-L/14 model(Radford et al., 2021), which was kept frozen to ensure computational efficiency. To model the temporal continuity inherent to sign language, we adopted the Sign Adapter module introduced in GFSLT (Zhou et al., 2023), which captures dependencies across consecutive frames. The Sign Adapter produces sign tokens via average pooling over temporally aligned features. These sign tokens serve as inputs for contrastive learning, which is performed using the AdamW optimizer with a learning rate=0.0001, β=(0.9,0.98), and weight decay=0.01. Training is performed for 256 epochs on the DGS dataset and 200 epochs on the KSL dataset.

Stage 2 For DGS translation, we employed Flan-T5-XL[3](Chung et al., 2024), a multilingual instruction-following model with strong capabilities in translation and text generation. In the case of KSL, we used pko-Flan-T5-Large[4], which shares the same model architecture but is pretrained on Korean corpora, due to Flan-T5-XL's limited proficiency in Korean. To preserve the pretrained linguistic knowledge of the language models, we applied Low-Rank Adaptation (LoRA) (Hu et al., 2022) during training, allowing efficient fine-tuning with minimal updates to the original parameters. LoRA parameters are set as follows: rank = 16, α = 32, target modules = q, v, and dropout = 0.1. Optimization is again conducted using AdamW(Loshchilov and Hutter, 2017) with the same configuration as in Stage 1. To integrate contrastive learning into this stage, we scale the contrastive loss by $\alpha = 0.1$ and add it to the cross-entropy loss.

A.2 Computing Environment

All experiments were conducted on a single NVIDIA A6000 (49GB) GPU with CUDA 12.3 and PyTorch 2.0.1. For dataset-specific configurations, DGS experiments used a batch size of 256 **(Stage 1)** and 4 **(Stage 2)**, while KSL used 32 and 8.

[3]https://huggingface.co/google/flan-t5-xl

[4]https://huggingface.co/paust/pko-flan-t5-large

A.3 Prompt Construction

The input fed to the LLM follows a unified structure across both DGS and KSL, formatted as a sign tokens followed by an instruction. For each instance, two translation pairs are randomly selected from a predefined multilingual pool to construct the retrieval-based exemplars. For DGS, the candidate languages are French, Spanish, and English; for KSL, they are Chinese, Japanese, and English. The final prompt format is structured as follows Table 4, and example is Table 5:

```
[VIDEO] Instruction
Retrieved Example (Random Pair 1) = DE/KO Translation
Retrieved Example (Random Pair 2) = DE/KO Translation
```

Table 4: Format of LLM Input

Sign Video Input:	[VIDEO]
Instruction:	Translate the given sentence into German.
In Context Exemplars:	et maintenant les prévisions météo pour demain, jeudi 12 août= und nun die wettervorhersage für morgen donnerstag den zwölften august and now the weather forecast for tomorrow, Thursday the twelfth of August= und nun die wettervorhersage für morgen donnerstag den zwölften august

Table 5: An example of DGS prompt used in this paper.

B Data Distribution

Dataset	Domain	Train	Dev	Test	Avg. Frame	Vocab Size
DGS	Weather	7,096	519	642	116	3K
KSL	Total	59,846	7,470	7,466	176	4K
	Healthcare	3,756	493	504	183	–
	Tourism	16,540	2,063	2,009	180	–
	Public Services	22,595	2,694	2,819	175	–
	Shopping	16,955	2,220	2,134	170	–

Table 6: Statistics of the datasets used in our experiments. DGS comprises weather domain, while KSL spans four domains with broader linguistic and contextual diversity.

Overview We evaluate our method on both Korean and German Sign Language datasets. Table 6 summarizes the datasets used in our experiments. To evaluate cross-linguistic and cross-domain generalization in gloss-free SLT, we consider two sign language corpora: KSL and DGS.

KSL The KSL dataset is a large-scale, multi-domain corpus released by the National Institute of Korean Language[5]. It contains a total of 74,782 sentence-aligned sign videos, partitioned into 59,846 for training, 7,470 for validation, and 7,466 for testing. The dataset covers four distinct domains—Tourism, Public Services, Shopping, and Healthcare—providing a broad linguistic

[5]https://www.korean.go.kr/

and contextual range for evaluating domain generalization.

DGS For DGS, we use the RWTH-PHOENIX-Weather 2014T dataset (Camgoz et al., 2018), a widely used benchmark in sign language translation. This dataset consists of 8,257 video–text pairs (7,096 training, 519 validation, 642 test), all sourced from televised weather broadcasts.

C Qualitative Example

Golden	아이들이 갑자기 **소변이 마렵다**고 해서요 (The children suddenly said they **needed to pee**.)
Zero	어서 오십시오 유행이 돼서 그런가 봐요 (Welcome. I guess it's because it's become a trend.)
Rand	네 아이들이 갑자기 고장이 나고 싶어서요 (Yes, the children suddenly said they wanted to break down.)
Sim	아이가 갑자기 **화장실을 가고 싶다**고 해서요 (A child suddenly said they **wanted to go to the bathroom**.)
Golden	객실 내에서 **흡연**이 가능한가요? (Is **smoking** allowed in the room?)
Zero	객실 내에서 통화가 가능한가요? (Is making a phone call allowed in the room?)
Rand	아 그래요 객실 내에서 말하기가 가능한가요 (Oh, really. Is speaking allowed in the room?)
Sim	객실 내에서 **흡연**이 가능한가요? (Is **smoking** allowed in the room?)
Golden	코스 소요시간은 약 **1시간** 정도 걸립니다 (The course takes about **one hour**.)
Zero	장소마다 소요시간은 약 3시간 정도 소요됩니다 (Each place takes about three hours.)
Rand	현장 소요시간은 약 **1시간** 정도 소요됩니다 (On-site time takes about **one hour**.)
Sim	장소까지의 소요시간은 약 **1시간** 정도 걸립니다 (Travel time to the place takes about **one hour**.)

Table 7: Qualitative examples grouped by reference and similarity level in KSL.

Qualitative Examples Table 7 presents qualitative examples from the KSL dataset, categorized by reference type and retrieval similarity level. Each block illustrates the target reference sentence **Golden**, followed by three retrieved examples: **Zero**, **Rand**, and **Sim (Ours)**. These examples demonstrate that semantically aligned prompts (Sim) tend to preserve contextual and domain-specific information closely aligned with the gold reference. In contrast, Zero and Rand examples often diverge in topic or omit key semantic elements, which may hinder accurate LLM-based translation. This comparison underscores the importance of semantic relevance in prompt design.