Selective Contrastive Learning For Gloss Free Sign Language Translation

Changhao Lai[1,2,3*], **Rui Zhao**[1,2,3*], **Xuewen Zhong**[1,2,3],
Jinsong Su[1,2] and **Yidong Chen**[1,2,3†]
[1]School of Informatics, Xiamen University, China
[2]Key Lab of Digital Protection and Intelligent Processing of Intangible
Cultural Heritage of Fujian-Taiwan (XMU), Ministry of Culture and Tourism, China
[3]National Language Resources Monitoring and
Research Center for Education and Teaching Media, Xiamen University, China
laichanghao@stu.xmu.edu.cn ydchen@xmu.edu.cn

Abstract

Sign language translation (SLT) converts continuous sign videos into spoken-language text, yet it remains challenging due to the intrinsic modality mismatch between visual signs and written text, particularly in gloss-free settings. Recent SLT systems increasingly adopt CLIP-like Vision-Language pretraining (VLP) for cross-modal alignment, but the random in-batch contrast provides few, batch-dependent negatives and may mislabel semantically similar (or even identical) pairs as negatives, introducing noisy and potentially inconsistent alignment supervision. In this work, we first conduct a preliminary trajectory-based analysis that tracks negative video-text similarity over training. The results show that only a small subset of negatives exhibits the desired behavior of being consistently pushed away, while the remaining negatives display heterogeneous and often non-decreasing similarity dynamics, suggesting that random in-batch negatives are frequently uninformative for effective alignment. Inspired by this, we propose Selective Contrastive Learning for SLT (SCL-SLT) with a Pair Selection (PS) strategy. PS scores candidate negatives using similarity dynamics from reference checkpoints and constructs mini-batches via a curriculum that progressively emphasizes more challenging negatives, thereby strengthening contrastive supervision while reducing the influence of noisy or semantically invalid negatives.

1 Introduction

Sign language serves as a primary means of communication for the Deaf and hard-of-hearing community worldwide. As a vision-centric language, it possesses the full range of fundamental linguistic properties, with meaning expressed through the coordinated use of manual and non-manual articulatory cues, e.g., hand shapes, movements, and facial expressions. These characteristics present significant challenges for automatic sign language understanding and have simultaneously drawn growing attention from the research community, particularly in sign language translation (SLT), which aims to translate continuous sign video sequences into spoken language text (Camgoz et al., 2018, 2020; Zhou et al., 2021; Chen et al., 2022; Zhou et al., 2023a; Fu et al., 2024; Zhao et al., 2024; Kim et al., 2025; Zhang et al., 2025; Fu et al., 2025).

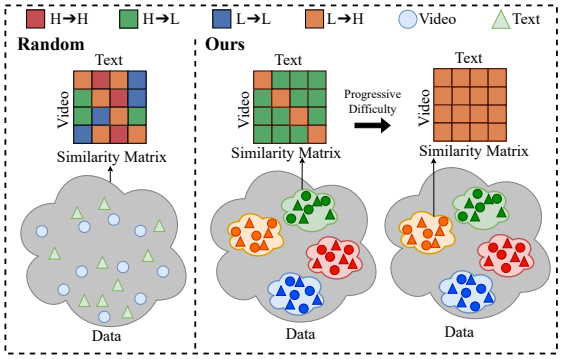


Figure 1: Comparison between the vanilla contrastive learning (left) and our Selective Contrastive Learning (right). SCL progressively selects highly informative negatives for efficient and effective alignment in SLT.

However, SLT inherently faces a substantial modality mismatch between sign videos and written text, which remains a key bottleneck for translation quality and keeps current performance suboptimal. This challenge is further amplified in gloss-free[1] SLT (GFSLT), where gloss annotations are unavailable, and the model is forced to learn the video-to-text mapping without an intermediate supervision for alignment. To mitigate this gap, a prominent line of work adopts CLIP-like

*Equal contribution.
†Corresponding Author.

[1]Glosses are word-level spoken-language annotations that roughly correspond to sign meanings, typically transcribed in a fixed sign-by-sign order.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 45648–45660
July 2-7, 2026

Figure 2: Semantically similar or identical instances.

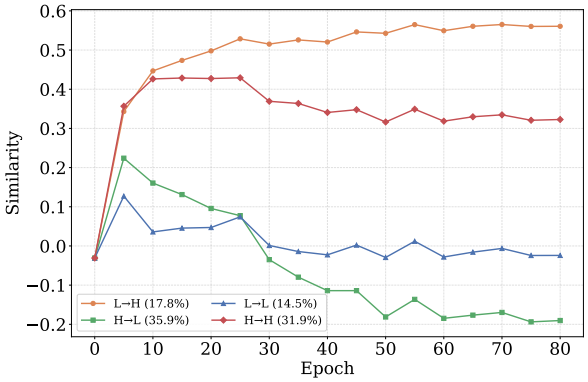


Figure 3: Similarity trajectories of four categories.

Vision-Language pretraining (VLP) to strengthen cross-modal alignment, yielding consistent empirical gains in recent SLT systems (Zhou et al., 2023a; Cheng et al., 2023; Ye et al., 2024; Liang et al., 2024; Chen et al., 2025). As illustrated in Figure 1 (left), these methods learn a shared embedding space through mini-batch contrast, pulling each matched video–text pair closer while pushing it away from all other non-matching pairs within the same mini-batch. Despite its effectiveness, this in-batch design provides each example with only a limited and batch-dependent set of negative pairs per update, and the problem is further exacerbated by the high computational cost associated with spatio-temporal video modeling, which often limits the achievable batch size. More critically, in-batch contrastive learning implicitly assumes that negatives are semantically distinct from the anchor. In practice, however, semantically similar or even semantically identical samples may be treated as negatives, as exemplified in Figure 2, potentially introducing conflicting supervision signals and impeding reliable cross-modal alignment.

To address this limitation, we conduct a preliminary study that tracks how the similarity of negative video-text pairs evolves over training (Section 2). The results in Figure 3 reveal that only a small fraction of negatives provide informative contrastive supervision, in the sense that their similarity is consistently reduced over training (H $\rightarrow$ L). Most negatives instead fall into two broad regimes: *easy* pairs that remain consistently dissimilar (L $\rightarrow$ L), and *hard* pairs that stay highly similar with substantial fluctuations (H $\rightarrow$ H), while a smaller portion exhibits increasingly challenging similarity over training (L $\rightarrow$ H). Building on these observations, we propose **S**elective **C**ontrastive **L**earning for SLT (**SCL-SLT**) with a Pair Selection (PS) strategy (Section 3.1). As shown in Figure 1 (right), PS dynamically prioritizes informative negative pairs to strengthen the contrastive supervision signal. The overall pipeline (Figure 5) consists of three stages. We first train a standard contrastive baseline with in-batch random negatives as a reference model. Then, checkpoints of the reference model are collected during reference training to score candidate negatives by their similarity trajectories (e.g., the change in similarity from early to late training), and retain highly informative negatives accordingly. Finally, we adopt a curriculum strategy to progressively shift the selected negatives from *easy* to *hard* during SLT training, moving from negatives that are already low-similarity or readily separable to those that remain highly similar or become increasingly similar.

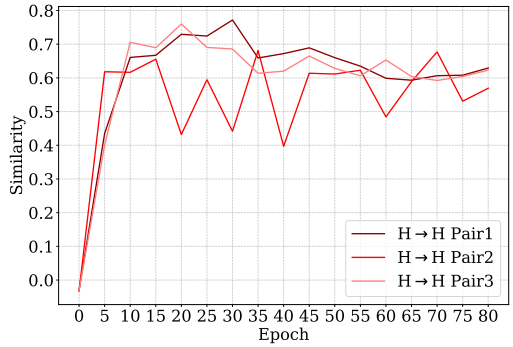
Unlike recent methods that heavily rely on massive Language Models (LLMs) or complex auxiliary annotations, SCL-SLT takes an orthogonal, data-centric approach. We demonstrate that rigorously optimizing intrinsic contrastive signals alone is fully sufficient to achieve superior translation precision, without introducing extraneous computational overhead or external data dependencies.

In summary, our contributions are as follows:

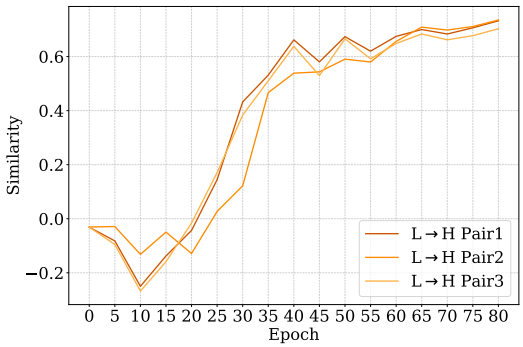
(1) We present a similarity trajectory analysis of in-batch negative pairs for CLIP-like contrastive learning in SLT, demonstrating that negatives contribute unevenly and impede reliable alignment.

(2) We propose SCL-SLT, which prioritizes informative negatives while mitigating the impact of noisy or semantically invalid negatives via a curriculum-guided Pair Selection mechanism.

(3) Extensive experiments on PHOENIX14T and CSL-Daily demonstrate that SCL-SLT significantly outperforms vanilla contrastive learning.



(a) Example of negatives in H → H.



(b) Example of negatives in L → H.

Figure 4: (a) The negative pairs in H → H exhibit increasing similarity over training, (b) The negatives in L → H show high similarity with fluctuations. Both cases demonstrate resistance to distinction during contrastive learning.

2 Preliminary

For an SLT dataset $D = \{V_i, T_i\}_{i=1}^{N}$ containing N paired sign video-text instances, the goal is to translate each source video V_i into its target sentence T_i. Due to the inherent nature of sign language and SLT maps sign videos to written text, the modality mismatch between the two remains a key challenge and often limits translation quality. Accordingly, a growing body of work adopts CLIP-like VLP to strengthen video-text alignment.

Limitations of In-Batch Negatives. Given a mini-batch $\mathcal{B} = \{V_i, T_i\}_{i=1}^{B}$ sampled from the training data, CLIP-like training treats each paired (V_i, T_i) as a positive, while pairing V_i with other texts $T_j (i \neq j)$ yields in-batch negatives. The resulting similarity matrix $S \in \mathbb{R}^{B\times B}$ places positives on the diagonal and negatives on the off-diagonal entries. The contrastive objective then pulls matched pairs together and pushes mismatched pairs apart in a shared embedding space (Figure 1, right). Under this in-batch design, each video is contrasted against only $B-1$ negatives per update, and the participating negatives depend on the random mini-batch composition. Consequently, the per-update negative coverage is $(B-1)/(N-1)$, which can be very small when $B \ll N$, thus many potentially *informative negatives are seldom observed in typical training.* This issue is further exacerbated by the high computational cost of spatio-temporal video modeling, which often necessitates smaller batch sizes and consequently limits the number and diversity of in-batch negatives. More importantly, the common assumption that *in-batch negatives are semantically distinct may not hold in SLT corpora*: the PHOENIX14 (Camgoz et al., 2018) includes many pairs with substantially similar semantics, and CSL-Daily (Zhou et al., 2021) further contains numerous pairs with identical target sentences (Figure 2). Therefore, semantically similar or even identical pairs can be treated as negatives (false negatives), producing inconsistent supervision signals that can hinder robust cross-modal alignment. These observations motivate a more principled treatment of negatives that accounts for both coverage and semantic validity, which is the focus of our approach.

Quantitative Study of In-Batch Negatives. To investigate how negative video-text similarities evolve over training, we train a contrastive model on the PHOENIX14T following the GFSLT-VLP framework (Zhou et al., 2023a) and track the similarity changes of each paired sample at intervals of 5 epochs. A clear pattern emerges from the similarity trajectories. while the diagonal elements (positive) align perfectly with the goal of contrastive learning, the off-diagonal elements (negatives) exhibit substantial heterogeneity. It suggests that while positives are effectively pulled closer, *not all negatives are successfully pushed apart.* We further quantitatively analyze the changes in similarity of these negatives using a linear least-squares regression. Finally, we stratify these negatives into four distinct categories based on their fitted trends: H → L (decreasing similarity), L → H (increasing similarity), L → L (consistently low similarity), and H → H (consistently high similarity). As shown in Figure 3, only 35.9% of negatives follow the desired H → L pattern, suggesting that a limited portion of negatives are consistently pushed apart. The remaining negatives do not exhibit a

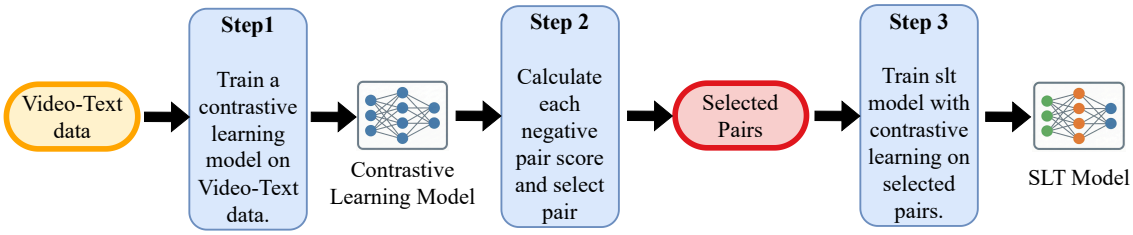


Figure 5: **Overview of the SCL-SLT pipeline.** The process consists of three stages: (Step 1) Training a preliminary contrastive learning model on video-text data, (Step 2) Computing similarity scores to select informative negative pairs, and (Step 3) Fine-tuning the target SLT model using contrastive learning on the selected pairs.

decreasing trend: 14.5% remain consistently low, 31.9% remain consistently high, and 17.8% unexpectedly increase over training. Representative cases in H $\rightarrow$ H and L $\rightarrow$ H subsets are shown in Figure 4, and detailed classification criteria are provided in Appendix A.

Overall, these observations suggest that negative pairs contribute unevenly to learning, and prioritizing informative negatives with high contrastive utility is crucial for improving cross-modal alignment. To this end, we introduce a Pair Selection strategy that systematically filters candidate negatives with high contrastive utility.

3 Method

As illustrated in Figure 5, we first train a reference video-text contrastive model (Step 1, in Section 2). We then apply the proposed Pair Selection strategy to construct mini-batches with informative and semantically valid negatives (Step 2, in Section 3.1). Finally, we present the selective contrastive training framework and its fine-tuning in SLT (Step 3, in Section 3.2).

3.1 Pair Selection

Inspired by the preliminary findings, we propose a Pair Selection mechanism that improves alignment efficiency by prioritizing informative negatives while mitigating the impact of noisy or semantically invalid negatives. We adopt a curriculum strategy (Bengio et al., 2009) that progressively shifts the selected negatives from *easy* pairs to *hard* pairs, corresponding to increasingly challenging similarity dynamics (Figure 6, top).

Trajectory-based difficulty proxy. We quantify negative-pair difficulty using the change in similarity over training. Let $\delta_{i,j}$ denote the *approximate* similarity change:

$$\delta_{i,j} = \hat{s}^K(V_i, T_j) - \hat{s}^0(V_i, T_j), \quad (1)$$

Algorithm 1 Curriculum-based Pair Selection

Input: dataset $\mathcal{D} = \{(V_i, T_i)\}_{i=1}^N$, curriculum ratio $\alpha \in [0, 1]$, batch size B
Output: batch list $\mathbb{B}$
$\mathcal{D}' \leftarrow \mathcal{D}$; $\mathbb{B} \leftarrow \emptyset$
while $\mathcal{D}' \neq \emptyset$ **do**
  sample $s_0 \sim \text{Uniform}(\mathcal{D}')$
  $\mathcal{C} \leftarrow \{s_0\}$; $\mathcal{D}' \leftarrow \mathcal{D}' \setminus \{s_0\}$
  while $|\mathcal{C}| < B$ **and** $\mathcal{D}' \neq \emptyset$ **do**
    $s \leftarrow \textsc{SelectByScore}(\mathcal{D}', \mathcal{C}, \alpha)$
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{s\}$; $\mathcal{D}' \leftarrow \mathcal{D}' \setminus \{s\}$
  end while
  $\mathbb{B} \leftarrow \mathbb{B} \cup \{\mathcal{C}\}$
end while
return $\mathbb{B}$

where $\hat{s}^k(V_i, T_j)$ denotes the *approximate* similarity between the i-th sign video and the j-th text computed with the k-th checkpoint of the reference contrastive model (Step 1 in Figure 5), and K denotes the final checkpoint. We define the batch score as the aggregate summation over all negative pairs within a batch:

$$\text{Score}_{\mathcal{C}} = \sum_{i=1}^{B} \sum_{j=1, j\neq i}^{B} \delta_{i,j}, \quad (2)$$

and aim to construct mini-batches whose scores follow a curriculum ratio $\alpha \in [0, 1]$. We adopt a linear schedule $\alpha = e/E$, where e denotes the current epoch and E denotes the total number of epochs in the selective contrastive training stage. A formal definition of $\delta_{i,j}$ is provided in Appendix A.

Relation to trajectory categories. Our curriculum is driven by the signed similarity change $\delta_{i,j}$ (Equation 1), which implicitly orders negative pairs by their trajectory patterns. Following the taxonomy in Section 2, we refer to *easy* negatives as those that are either pushed away over training

(H $\rightarrow$ L, decreasing similarity) or remain consistently dissimilar (L $\rightarrow$ L, consistently low similarity), and *hard* negatives as those that stay highly similar (H $\rightarrow$ H, consistently high similarity) or become more similar over training (L $\rightarrow$ H, increasing similarity). Since $\delta_{i,j}$ can be negative for L $\rightarrow$ H pairs and positive for H $\rightarrow$ L pairs, selecting negatives by increasing score naturally yields a curriculum that starts from decreasing-similarity negatives and gradually shifts towards increasing-similarity negatives as α increases. Pairs in the L $\rightarrow$ L and H $\rightarrow$ H regimes typically exhibit small $|\delta_{i,j}|$ magnitudes; thus, while they are not explicitly excluded by our selection rule, they tend to be less influential under a trajectory-change-driven curriculum.

Incremental score for greedy batch construction. Given a training dataset $\mathcal{D} = \{(V_i, T_i)\}_{i=1}^{N}$, exhaustively searching over all possible batch constructions to optimize the global score is computationally intractable due to combinatorial explosion and memory overhead, especially when $B \ll N$. We therefore cast negative selection as a batch construction problem and adopt a greedy procedure (Algorithm 1). Specifically, given a partially constructed batch $\mathcal{C}$ and a candidate pair $s_u = (V_u, T_u) \in \mathcal{D}'$, we define its *incremental score* as the score increase incurred by adding s_u into $\mathcal{C}$:

$$\begin{aligned}\Delta(s_u; \mathcal{C}) &= \text{Score}_{\mathcal{C}\cup\{s_u\}} - \text{Score}_{\mathcal{C}} \\ &= \sum_{(V_i,T_i)\in\mathcal{C}} (\delta_{i,u} + \delta_{u,i}),\end{aligned} \tag{3}$$

where $\delta_{i,u}$ measures the approximate similarity change between V_i and T_u, and $\delta_{u,i}$ is defined analogously between V_u and T_i (Equation 1). Accordingly, SELECTBYSCORE$(\mathcal{D}', \mathcal{C}, \alpha)$ computes $\Delta(s_u; \mathcal{C})$ for all $s_u \in \mathcal{D}'$, sorts candidates by Δ in ascending order, and selects the candidate at the α-quantile rank, i.e., $r = \lfloor \alpha \cdot (|\mathcal{D}'| - 1) \rfloor$. Concretely, we seed each batch with one randomly sampled pair from $\mathcal{D}'$, and then iteratively add the selected candidate until the batch reaches the target size B.

3.2 Selective Contrastive Learning

Our SCL-SLT framework consists of four components: a Sign Embedding module, a Visual Encoder, a Text Encoder, and a Decoder, as illustrated in Figure 6. We first perform selective contrastive training with PS to learn better-aligned video–text representations, while jointly optimizing the translation objective to preserve generation capability. We then conduct task-specific SLT fine-tuning, where the text branch used for alignment is detached and the model is trained with the translation loss only.

Sign Embedding. The Sign Embedding module converts an input sign video clip into a temporal feature sequence. It comprises a ResNet (He et al., 2016) backbone for spatial feature extraction, followed by two temporal blocks, each consisting of stacked 1D Convolution, Batch Normalization, and ReLU activation (Conv1D/BN/ReLU).

Visiual & Textual Encoders. We adopt a dual-encoder design for video-text representation learning. The Visual Encoder is a Transformer encoder initialized from a pre-trained sequence-to-sequence language model mBART-large-50 (Liu et al., 2020) and adapted with LoRA (Hu et al., 2022), enabling it to extract high-level visual representations from the Sign Embedding. The Text Encoder, also a Transformer encoder initialized from the same mBART-large-50 module, is kept frozen to provide robust linguistic priors for video-text alignment during the contrastive learning stage.

Video–Text Alignment. We adopt a CLIP-style contrastive objective to align the visual and textual feature spaces by increasing similarity for matched pairs and decreasing it for mismatched pairs. For sequence-level alignment, we consider three aggregation strategies to obtain global representations: (1) CLS pooling (Zhou et al., 2023a), (2) mean pooling over the feature sequences (Kim et al., 2025; Liang et al., 2024), and (3) the fine-grained CiCo aggregation protocol (Cheng et al., 2023). A comparative analysis is provided in Section 4.4. In our final implementation, the similarity matrices S_{V2T} and S_{T2V} are computed following CiCo, and the contrastive loss is computed as:

$$\begin{aligned}\mathcal{L}_{\text{CL}} = -\frac{1}{2B}\Big(&\sum_{i=1}^{B} \log \frac{\exp(S_{V2T}^{i,i}/\tau)}{\sum_{k=1}^{B}\exp(S_{V2T}^{i,k}/\tau)} \\ &+ \sum_{i=1}^{B} \log \frac{\exp(S_{T2V}^{i,i}/\tau)}{\sum_{k=1}^{B}\exp(S_{T2V}^{i,k}/\tau)}\Big),\end{aligned} \tag{4}$$

where B is the batch size and τ is a trainable temperature parameter. During selective contrastive training, we primarily optimize the contrastive alignment loss $\mathcal{L}_{\text{CL}}$. In addition, we retain the standard translation training signal and set its weight to 1.0 to maintain generation capability while learning aligned representations in all experiments.

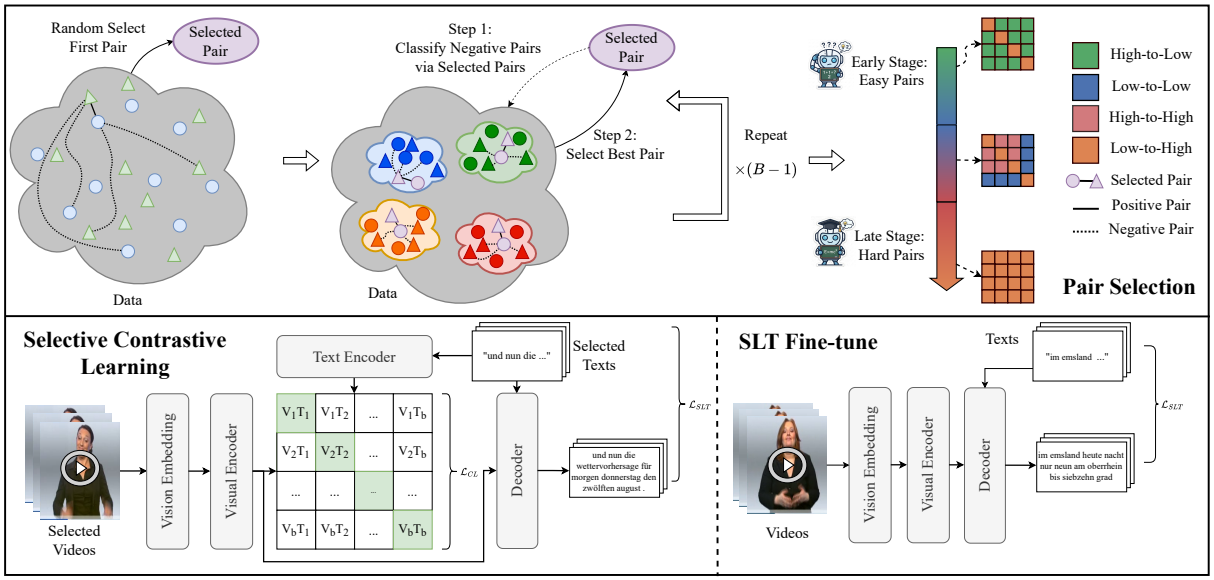


Figure 6: **Overview of the proposed SCL-SLT framework. (Top)** Illustration of the Pair Selection strategy. During batch construction, we initialize with a random positive pair and iteratively select subsequent pairs by evaluating candidates against the current selection. The process follows a curriculum learning (Bengio et al., 2009) schedule, progressively transitioning from "Easy Pairs" in early stages to "Hard Pairs" in later stages. **(Bottom)** In the first stage, **Selective Contrastive Learning (SCL)** utilizes selected pairs to align visual and textual modalities. In the second stage, **SLT Fine-tuning**, the auxiliary Text Encoder and alignment module are detached, allowing the model to focus exclusively on optimizing translation performance.

SLT Fine-tuning. The Decoder is a LoRA-adapted pre-trained seq2seq decoder. Conditioned on the sign video V and previously generated tokens $y_{<t}$, it generates the target sentence autoregressively. In the SLT fine-tuning stage, we detach the text branch used for alignment and optimize the translation objective only:

$$\mathcal{L}_{\text{SLT}} = -\sum_{t=1}^{|y|} \log P(y_t \mid V, y_{<t}). \tag{5}$$

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets. We evaluate our method on two benchmarks: **PHOENIX14T** (Camgoz et al., 2018), a German Sign Language dataset in the weather domain, and **CSL-Daily** (Zhou et al., 2021), a large-scale Chinese dataset covering daily topics. PHOENIX14T contains 7,096/519/642 pairs (Train/Dev/Test) with a vocabulary of 2,887. CSL-Daily, noted for its multi-signer diversity, comprises 18,401/1,077/1,076 samples with a vocabulary size of 2,343.

Evaluation Metrics. Following previous studies (Kim et al., 2025; Zhou et al., 2023a), we use BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) to measure the translation performance.

4.2 Implementation Details

Data Processing. We apply $4\times$ temporal downsampling, selecting a random frame per clip during training and the first frame during inference. Spatially, frames are resized to 256×256 and then cropped to 224×224, using random cropping for training and center cropping for inference.

Model Architecture. For sign embedding, we employ a ResNet-18 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009), followed by a temporal block (Conv1D-BN-Relu). The Visual Encoder, Text Encoder, and Decoder are initialized with the pre-trained `mbart-large-50` (Tang et al., 2020) ($L = 12, d_{model} = 1024, d_{ffn} = 4096$, 16 heads). During training, the Text Encoder is frozen, while the Visual Encoder and Decoder are fine-tuned via Low-Rank Adaptation (LoRA) with $r = 16$ and $\alpha = 32$.

Hyperparameters. We optimize both stages using AdamW (Loshchilov and Hutter, 2017) with 0.2 label smoothing, and batch sizes of 16 and 8 for Stages 1 and 2. The initial learning rate is 1×10^{-4} with a cosine decay (Loshchilov and Hutter, 2016). The training duration is set to 80 epochs for Stage 1 and 200 epochs for Stage 2. During inference, we utilize a beam size of 8.

Method	PHOENIX14T					CSL-Daily				
	R	B1	B2	B3	B4	R	B1	B2	B3	B4
GFSLT w/o VLP										
NSLT (2018)	29.70	27.10	15.61	10.82	8.35	-	-	-	-	-
NSLT+Bahdanau (2018; 2014)	31.80	32.24	19.03	12.83	9.58	-	-	-	-	-
NSLT+Luong (2018; 2015)	30.70	29.86	17.52	11.96	9.00	34.54	34.16	19.57	11.84	7.56
SLRT* (2020)	31.10	30.88	18.57	13.12	10.19	19.67	20.00	9.11	4.93	3.03
MMTLB† (2022)	38.60	40.57	26.99	19.58	15.18	26.70	27.22	15.90	10.61	7.61
GASLT (2023)	30.86	39.07	26.74	21.86	15.74	20.35	19.90	9.94	5.98	4.07
GFSLT (2023a)	40.93	41.39	31.00	24.20	19.66	35.16	37.69	23.28	14.93	9.88
Sign2GPT (2024)	48.90	49.54	35.96	28.83	22.52	42.36	41.75	28.73	20.60	15.40
FLa-LLM (2024)	45.27	46.29	35.33	28.03	23.09	37.25	37.13	25.12	18.38	14.20
SignLLM (2024)	44.49	45.21	34.78	28.05	23.40	39.91	39.55	28.13	20.07	15.75
SCL-SLT(Ours)	46.33	48.00	37.36	30.23	25.30	48.53	50.36	36.88	27.69	21.41
GFSLT w/ VLP										
GFSLT-VLP (2023a)	42.49	43.71	33.18	26.11	21.44	36.44	39.37	24.93	16.26	11.00
GFSLT-VLP-SignCL (2024)	49.04	49.76	36.85	29.97	22.74	48.92	47.47	32.53	22.62	16.16
LLAVA-SLT (2024)	50.44	51.20	37.51	29.39	23.43	**51.26**	52.15	36.24	26.47	20.42
C^{2}RL (2025)	**50.96**	**52.81**	**40.20**	**32.20**	**26.75**	48.21	49.32	36.28	27.54	21.61
SpaMo (2025)	46.57	49.80	37.32	29.50	24.32	47.46	48.90	36.90	26.78	20.55
MMSLT (2025)	47.97	48.92	38.12	30.79	25.73	48.92	49.87	36.37	27.29	21.11
SCL-SLT(Ours)	47.02	48.72	38.19	31.04	26.00	51.08	**52.81**	**39.28**	**29.82**	**23.25**

Table 1: Performance comparison with state-of-the-art gloss-free SLT methods on the PHOENIX14T and CSL-Daily sets. R and Bn denote ROUGE and BLEU-n, respectively. † indicates our reproduction under the gloss-free setting. For SLRT*, results are sourced from Yin et al. (2023) for PHOENIX14T and Zhou et al. (2023a) for CSL-Daily. The best results are highlighted in **bold**.

4.3 Main Results

Table 1 presents a quantitative comparison with state-of-the-art gloss-free SLT methods. Our SCL-SLT demonstrates strong overall performance, establishing new state-of-the-art records on the CSL-Daily dataset and achieving highly competitive results on PHOENIX14T.

Results on PHOENIX14T. SCL-SLT demonstrates highly competitive performance on the PHOENIX14T benchmark. Under the *w/o* VLP setting, it establishes a new state-of-the-art with a BLEU-4 of **25.30**. With VLP, SCL-SLT reaches **26.00**. While C^{2}RL (Chen et al., 2025) achieves 26.75 via auxiliary tasks, our approach yields comparable results through a fundamentally streamlined paradigm: maximizing intrinsic data utility via refined negative sampling. Furthermore, SCL-SLT surpasses the strong LLM-centric LLAVA-SLT (Liang et al., 2024) by **2.57** in BLEU-4 (26.00 vs. 23.43).

Results on CSL-Daily. On the more challenging CSL-Daily benchmark, our method achieves a remarkable **23.25 BLEU-4** with a substantial margin of **2.14** over the previous best C^{2}RL (21.61). While the LLM-based method LLAVA-SLT remains competitive in ROUGE (51.26 vs. 51.08), our approach significantly outperforms it in all precision-oriented BLEU metrics (e.g., +2.83 in BLEU-4).

Superiority over Methods *w/o* VLP. SCL-SLT achieves state-of-the-art performance on both PHOENIX14T and CSL-Daily benchmarks under the *w/o* VLP setting. Notably, SCL-SLT achieves a BLEU-4 score of **25.30** on PHOENIX14T and **21.41** on CSL-Daily, surpassing the strongest competitor (SignLLM) by substantial margins of **1.90** and **5.66**, respectively.

Superiority over Baselines *w/* VLP. A critical advantage of SCL-SLT lies in its elegant reliance on intrinsic data signals. While recent strong baselines depend on external annotations (e.g., frame-level motion descriptions in MMSLT (Kim et al., 2025)) or complex auxiliary objectives (e.g., Explicit Context Learning in C^{2}RL (Chen et al., 2025)), SCL-SLT avoids such architectural overhead. Instead, it employs a refined negative sampling strategy to systematically mine highly informative pairs directly from the existing dataset. Our empirical results demonstrate that maximizing data utility to establish robust cross-modal alignment is sufficient to achieve superior performance, entirely eliminating the need for extraneous complexity.

Method	PHOENIX14T		CSL-Daily	
	R	B4	R	B4
BaseLine(End-to-End)	41.81	21.97	41.04	16.31
w/ CL	43.55	22.03	47.77	20.59
w/ SCL	46.33	25.30	48.53	21.41
CL-SLT	46.13	25.01	48.34	20.70
SCL-SLT (Ours)	47.02	26.00	51.08	23.25

Table 2: Ablation study on the effectiveness of the SCL strategy. (Top) End-to-End SLT. (Bottom) Vision-Language Pretraining SLT.

Dataset	Number of Videos for Unique Text			
	1	2	3	≥ 4
PHOENIX14T	6,811	21	5	16
CSL-Daily	45	1,480	4,850	203

Table 3: **Statistics of target redundancy in SLT benchmarks.** We report the number of unique target sentences associated with N different sign videos. Notably, CSL-Daily exhibits a high degree of redundancy, with most texts corresponding to multiple videos (e.g., $N = 3$).

4.4 Ablation Studies

In this section, we conduct ablation studies to validate the effectiveness of the proposed method. Unless otherwise specified, all ablation experiments are performed on the PHOENIX14T test set.

Effectiveness of SCL. Table 2 reveals that *w/* CL works well on the multi-topic CSL-Daily (+4.28) but fails on the weather-focused PHOENIX14T (+0.06), as random sampling lacks the granularity to distinguish similar intra-domain samples. *w/* SCL resolves this by filtering for informative negatives, boosting PHOENIX14T performance by **2.56** over *w/* CL. Although the initial gain on CSL-Daily is modest (+0.82) due to sparse intra-topic samples, the fine-tuned SCL-SLT surpasses CL-SLT by a remarkable **2.55 BLEU-4**. This confirms that even when immediate end-to-end gains are limited, SCL captures superior structural features that are critical for maximizing final translation quality.

As Table 3 shows, text-to-video multiplicity introduces severe false negatives. Unlike the baseline *w/* CL, our *w/* SCL filters these identical-pair collisions. Beyond identical texts, high intra-dataset similarity exacerbates the false negative problem, directly amplifying the gains from our Pair Selection (PS). This is prominent on PHOENIX14T (*w/o* VLP), where the narrow weather domain yields highly homogeneous texts. Furthermore, even in broader datasets like CSL-Daily (*w/* VLP), inherent sign language vocabulary limits inevitably introduce structural similarities. By navigating these intrinsic similarities, PS learns robust cross-modal representations across diverse domains.

Method	R	B1	B2	B3	B4
Random	43.55	44.74	33.91	26.84	22.03
Hard-Only	28.94	31.54	21.46	15.90	12.41
Easy-Only	44.93	46.57	35.92	28.83	24.00
Linear	46.31	48.36	37.27	29.76	24.59
Sqrt	45.86	47.56	36.73	29.53	24.54
Log	46.33	48.00	37.36	30.23	25.30

Table 4: Ablation study on pair selection strategies.

Method	R	B1	B2	B3	B4
Mean Pooling	28.52	30.57	20.92	15.66	12.44
`[CLS]` Pooling	33.03	35.20	24.98	18.73	14.85
CiCo (2023)	46.33	48.00	37.36	30.23	25.30

Table 5: Ablation study on representation aggregations.

To further assess the influence of distinct contrastive strategies, subsequent ablation experiments are performed under the *w/* SCL setting.

Impact of Curriculum Scheduling Strategies. As shown in Table 4 To investigate the influence of the curriculum scheduling parameter α on translation performance, we evaluate both static and dynamic settings. Specifically, $\alpha = 0$ indicates that only the most easily distinguishable simple samples are selected (Easy-Only), while $\alpha = 1$ means that only highly challenging hard samples are selected (Hard-Only). Let k denote the current training epoch and E_{ref} represent the total number of epochs. In addition to the two extreme static strategies, we focus on comparing the following three dynamic scheduling functions, where α gradually transitions from 0 to 1 as training progresses: (1) **Linear Scheduling (Linear:** $\alpha = k/E_{\text{ref}}$**)** (2) **Logarithmic Scheduling (Log:** $\alpha = \ln\left(1 + (k/E_{\text{ref}}) \cdot (\mathrm{e} - 1)\right)$**)** (3) **Square Root Scheduling (Sqrt:** $\alpha = \sqrt{(k/E_{\text{ref}})}$**).**

Impact of Similarity Computation Methods. As show in Table 5.To investigate the influence of different similarity matrix calculation strategies, we conducted experiments using three approaches: the `[CLS]` Pooling, Mean Pooling, and CiCo (Cheng et al., 2023). As shown in Table 5, the CiCo method yields significantly superior performance, outperforming both the `[CLS]` Pooling and Mean Pooling baselines by a substantial margin.

Interval	R	B1	B2	B3	B4
1	45.22	47.64	36.91	29.66	24.68
5	**46.33**	**48.00**	**37.36**	**30.23**	**25.30**
10	15.89	16.33	10.26	7.73	6.26

Table 6: Ablation study on the checkpoint sampling interval for trajectory estimation.

Impact of Checkpoint Sampling Interval. Table 6 evaluates the sampling granularity for similarity trajectory estimation. An interval of 5 epochs yields the optimal BLEU-4 of **25.30**. Sampling too frequently (interval of 1) slightly degrades performance (24.68) by incorporating overly sensitive micro-fluctuations and training noise. Conversely, a sparse interval of 10 causes a catastrophic drop to 6.26 BLEU-4. This indicates that a coarse temporal resolution fails to capture the true dynamics of negative pairs, effectively misguiding the Pair Selection curriculum. Thus, an interval of 5 optimally balances trend capturing and noise smoothing.

5 Related Work

5.1 Gloss-Free SLT with Contrastive Learning

GFSLT-VLP (Zhou et al., 2023a) first introduced CLIP-based pre-training to SLT, aligning global video features (via `[CLS]`) with text. CiCo (Cheng et al., 2023) improved upon this by performing fine-grained frame-text alignment without the global token constraint. To further boost representation quality, MMSLT (Kim et al., 2025) and C^2RL (Chen et al., 2025) introduced auxiliary supervision signals, utilizing action descriptions and an ECL loss, respectively. LLaVA-SLT (Liang et al., 2024) combined contrastive alignment with Large Language Models via instruction tuning, achieving robust translation performance.

5.2 Gloss-Free SLT without Contrastive Learning

To alleviate the dependency on expensive gloss annotations, GASLT (Yin et al., 2023) proposes a gloss attention mechanism that implicitly captures gloss-level representations in a fully gloss-free setting. More recently, leveraging the powerful reasoning and generative capabilities of Large Language Models (LLMs), methods such as Sign2GPT (Wong et al., 2024) and SignLLM Gong et al. (2024) have integrated pre-trained LLMs into the SLT framework. These approaches facilitate robust, direct translation from continuous sign videos to spoken sentences by harnessing the linguistic priors of LLMs.

5.3 Data Selection Strategies

Recent advancements in NLP highlight that data quality often outweighs quantity (Zhou et al., 2023b), driving automated selection methods based on model scoring or training dynamics (Chen et al., 2023; Lin et al., 2024; Xia et al., 2024). Inspired by this paradigm, we extend the focus from sample-level filtering in general NLP to pair-level selection in cross-modal SLT contrastive learning.

6 Conclusion

In this work, we investigate the role of in-batch negatives in CLIP-like contrastive learning in SLT and find that negative pairs exhibit highly uneven training dynamics, leading to unreliable video-text alignment. Motivated by this, we proposed SCL-SLT with a curriculum-guided Pair Selection strategy that prioritizes informative negatives while reducing the influence of noisy or semantically invalid ones. Our SCL-SLT improves over the vanilla contrastive learning and achieves new state-of-the-art performance on PHOENIX14T and CSL-Daily.

Limitations

A limitation of our current framework is the reliance on a separately trained Contrastive Learning Model to score and select negative pairs. This additional training stage introduces extra computational overhead and inference latency during the data preparation phase. To mitigate this, future work could explore leveraging powerful off-the-shelf pre-trained language models to compute semantic similarity directly. Replacing the trained reference model with such open-source alternatives would significantly streamline the pipeline and reduce resource consumption.

Ethics Statement

Our SCL-SLT framework aims to assist communication for the Deaf and hard-of-hearing community, but it is an experimental exploration of methodology and is not yet ready for real-world production or deployment. Furthermore, as our model is trained on specific regional datasets (PHOENIX14T and CSL-Daily), it may inherit data biases and exhibit degraded performance on

out-of-distribution signers, diverse dialects, or unconstrained environments. We exclusively utilized public datasets and advocate for the responsible, privacy-preserving deployment of SLT technologies.

Acknowledgements

We are grateful for the efforts and time of the reviewers and the committee. This work was supported in part by the National Natural Science Foundation of China under Grant 62476232, Grant 62076211, and in part by First Batch of Projects for the 2025 "Intergovernmental International Science, Technology and Innovation Cooperation" of the National Key Research and Development Program of China under Grant 2025YFE0121700.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48.

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*, pages 10023–10033.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and 1 others. 2023. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

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*, pages 5120–5130.

Zhigang Chen, Benjia Zhou, Yiqing Huang, Jun Wan, Yibo Hu, Hailin Shi, Yanyan Liang, Zhen Lei, and Du Zhang. 2025. C^2RL: Content and context representation learning for gloss-free sign language translation and retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*.

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*.

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*, pages 19016–19026.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Biao Fu, Liang Zhang, Peigen Ye, Pei Yu, Cong Hu, Xiaodong Shi, and Yidong Chen. 2025. Improving end-to-end sign language translation via multi-level contrastive learning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:1230–1242.

Honghao Fu, Liang Zhang, Biao Fu, Rui Zhao, Jinsong Su, Xiaodong Shi, and Yidong Chen. 2024. Signer diversity-driven data augmentation for signer-independent sign language translation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2182–2193, Mexico City, Mexico. Association for Computational Linguistics.

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.

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, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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.

Jungeun Kim, Hyeongwoo Jeon, Jongseong Bae, and Ha Young Kim. 2025. Leveraging the power of mllms for gloss-free sign language translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21048–21058.

Han Liang, Chengyu Huang, Yuecheng Xu, Cheng Tang, Weicai Ye, Juze Zhang, Xin Chen, Jingyi Yu, and Lan Xu. 2024. Llava-slt: Visual language tuning for sign language translation. *arXiv preprint arXiv:2412.16524*.

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

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and 1 others. 2024. Rho-1: Not all tokens are what you need. *arXiv preprint arXiv:2404.07965*.

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*, 8:726–742.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

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

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

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.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

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*.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*.

Jinhui Ye, Xing Wang, Wenxiang Jiao, Junwei Liang, and Hui Xiong. 2024. Improving gloss-free sign language translation by reducing representation density. *Advances in Neural Information Processing Systems*, 37:107379–107402.

Aoxiong Yin, Tianyun Zhong, Li Tang, Weike Jin, Tao Jin, and Zhou Zhao. 2023. Gloss attention for gloss-free sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2551–2562.

Ruiquan Zhang, Rui Zhao, Zhicong Wu, Liang Zhang, Haoqi Zhang, and Yidong Chen. 2025. Dynamic feature fusion for sign language translation using hypernetworks. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6227–6239, Albuquerque, New Mexico. Association for Computational Linguistics.

Rui Zhao, Liang Zhang, Biao Fu, Cong Hu, Jinsong Su, and Yidong Chen. 2024. Conditional variational autoencoder for sign language translation with cross-modal alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19643–19651.

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

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023b. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

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*, pages 1316–1325.

A Classification of Negatives

To characterize how negative video–text similarities evolve during training, we train a reference contrastive model on PHOENIX14T following the GFSLT-VLP framework (Zhou et al., 2023a). We save checkpoints every 5 epochs over E_{ref} training epochs, yielding a checkpoint index set $\mathcal{K} = \{0, 5, \ldots, E_{\text{ref}}\}$, and denote the final checkpoint by $K = \max(\mathcal{K})$.

Similarity trajectories. For each checkpoint $k \in \mathcal{K}$, we compute similarities between all video-text pairs and collect negative similarities $\{s^k(V_i, T_j)\}_{k\in\mathcal{K}}$ for all $i \in \{1, \ldots, N\}$ and $j \neq i$:

$$s^k(V_i, T_j) = \frac{\text{VE}^k(V_i) \cdot \text{TE}^k(T_j)}{|\text{VE}^k(V_i)|_2 |\text{TE}^k(T_j)|_2} \quad (6)$$

where $\text{VE}^k(V_i)$ and $\text{TE}^k(T_j)$ denote the high-dimensional global feature vectors extracted by the

visual and text encoders for the source video V_i and text T_j at the k-th checkpoint, respectively. Here, $\|\cdot\|_2$ represents the L_2 norm (i.e., Euclidean length) of the vector, which applies L_2 normalization to the feature representations in the denominator. The numerator calculates the dot product of the two vectors. Through this normalization, the ratio is mathematically equivalent to the cosine of the angle between the feature vectors, stably bounding the resulting scalar similarity score within the $[-1, 1]$ interval.

We then fit each negative trajectory using linear least-squares regression:

$$\min_{a,b} \sum_{k \in \mathcal{K}} \left(s^k(V_i, T_j) - (ak + b)\right)^2, \quad (7)$$

which yields a smoothed approximation:

$$\hat{s}^k(V_i, T_j) = ak + b, \quad k \in \mathcal{K}. \quad (8)$$

High/Low partition and change magnitude. We define the mean final negative similarity as:

$$\hat{s}_{\text{mean}} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \hat{s}^K(V_i, T_j), \quad (9)$$

and label a negative pair as *high* (H) if $\hat{s}^K(V_i, T_j) > \hat{s}_{\text{mean}}$ and *low* (L) otherwise. The similarity change $\delta_{i,j}$ is computed as in Equation 1, and we use a threshold ϵ to decide whether the change is substantial. Here, we set $\epsilon = 0.2$.

Classification rules. Finally, each negative pair is categorized based on its final similarity level and change direction:

$$\begin{cases} \text{L} \to \text{L}, & \text{if } \hat{s}^K(V_i, T_j) \leq \hat{s}_{\text{mean}} \text{ and } |\delta_{i,j}| \leq \epsilon, \\ \text{H} \to \text{H}, & \text{if } \hat{s}^K(V_i, T_j) > \hat{s}_{\text{mean}} \text{ and } |\delta_{i,j}| \leq \epsilon, \\ \text{L} \to \text{H}, & \text{if } \hat{s}^K(V_i, T_j) > \hat{s}_{\text{mean}} \text{ and } \delta_{i,j} > \epsilon, \\ \text{H} \to \text{L}, & \text{if } \hat{s}^K(V_i, T_j) \leq \hat{s}_{\text{mean}} \text{ and } \delta_{i,j} < -\epsilon. \end{cases} \quad (10)$$

B Contrastive Learning Model Configuration.

To ensure architectural consistency between the Contrastive Learning Model (inference model) in Figure 5 and the downstream SLT Model, we utilize the identical Visual and Text Encoders described in Section 3.2. We replicated the preliminary analysis from Section 2 using this updated configuration, specifically a 12-layer encoder optimized via LoRA, consistent with Section 4. As illustrated in Figure 7, while the category proportions show minor deviations, the "well-behaved" negative samples (L → L and H → L) account for only 46.5%. Notably, the distribution of similarity scores exhibits a narrower range. This is attributed to the use of LoRA with a reduced learning rate (1×10^{-4}), which restricts the magnitude of feature updates to better preserve the LLM's linguistic capabilities. We employ this calibrated model to calculate batch scores for pair selection.

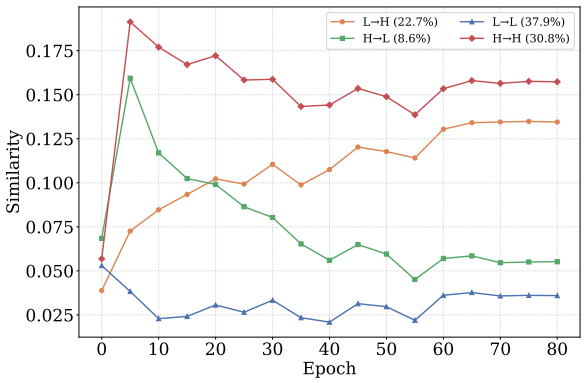


Figure 7: The average cosine similarity curves for negative pair categories based on Section 4 architecture

Method	PHOENIX14T				
	R	B1	B2	B3	B4
BaseLine(End-to-End)	37.38	38.37	28.52	22.27	18.16
GFSLT-VLP†	27.89	28.13	18.87	13.84	10.96
GFSLT-VLP(Ours)	40.67	41.82	31.49	24.72	20.27

Table 7: **Analysis of Contrastive Learning effectiveness.** We compare the baseline with GFSLT-VLP variants on PHOENIX14T. † denotes that the method was adapted to an end-to-end architecture for a fair comparison.

C Impact of Contrastive Learning on SLT

To investigate the impact of Contrastive Learning (CL) on SLT, we reproduced the GFSLT-VLP (Zhou et al., 2023a) method on the PHOENIX14T dataset. For a fair comparison with our baseline, we adapted GFSLT-VLP into an end-to-end architecture (denoted by †), where the visual encoder and translation decoder are trained jointly from scratch for 200 epochs.

As shown in Table 7, directly incorporating a contrastive objective into the end-to-end SLT baseline results in a precipitous performance drop (e.g., BLEU-4 falls from 18.16 to 10.96). This observation aligns with the findings in Section 4.4, where

adding standard CL yielded negligible gains (+0.06 BLEU-4). Conversely, when employing CL as a pre-training stage (as in the original GFSLT-VLP and our SCL-SLT), it consistently boosts translation performance (reaching 20.27 BLEU-4). This suggests that CL is most effective when used to learn robust representations prior to the translation task, rather than as a simultaneous auxiliary loss in an end-to-end pipeline.