Improving Gloss-free Sign Language Translation by Reducing Representation Density

Jinhui Ye[1] **Xing Wang**[4] **Wenxiang Jiao**[4,†] **Junwei Liang**[1,2,†] **Hui Xiong**[1,2,3,†]
[1]Artificial Intelligence Thrust, HKUST (Guangzhou), Guangzhou, China
[2]Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China
[3]Guangzhou HKUST Fok Ying Tung Research Institute [4]Tencent AI Lab
jye624@connect.hkust-gz.edu.cn; junweiliang@hkust-gz.edu.cn;
xionghui@ust.hk; {brightxwang,joelwxjiao}@tencent.com

Abstract

Gloss-free sign language translation (SLT) aims to develop well-performing SLT systems with no requirement for the costly gloss annotations, but currently still lags behind gloss-based approaches significantly. In this paper, we identify **a representation density problem** that could be a bottleneck in restricting the performance of gloss-free SLT. Specifically, the representation density problem describes that the visual representations of semantically distinct sign gestures tend to be closely packed together in feature space, which makes gloss-free methods struggle with distinguishing different sign gestures and suffer from a sharp performance drop. To address the representation density problem, we introduce a simple but effective contrastive learning strategy, namely SignCL, which encourages gloss-free models to learn more discriminative feature representation in a self-supervised manner. Our experiments demonstrate that the proposed SignCL can significantly reduce the representation density and improve performance across various translation frameworks. Specifically, SignCL achieves a significant improvement in BLEU score for the Sign Language Transformer and GFSLT-VLP on the CSL-Daily dataset by 39% and 46%, respectively, without any increase of model parameters. Compared to Sign2GPT, a state-of-the-art method based on large-scale pre-trained vision and language models, SignCL achieves better performance with only 35% of its parameters. Implementation and Checkpoints are available at https://github.com/JinhuiYE/SignCL.

1 Introduction

Sign languages are the primary form of communication for millions of deaf individuals. Sign language translation (SLT) aims to convert sign language into fluent spoken language sentences, which is a challenging task as it needs to extract information from continuous video and translate it into discrete text tokens. Most prior studies promoted the SLT by utilizing intermediate representations, namely gloss annotations, either directly or indirectly [3, 48, 53, 8, 49, 46, 41]. Gloss annotations are beneficial as they provide a simplified representation and sequential ordering of each gesture within continuous sign videos, which aids in representation learning for visual encoders. However, the creation of sign language translation datasets with gloss annotations is both resource-intensive and time-consuming.

Recently, there has been a shift towards gloss-free sign language translation methods, which do not rely on gloss annotations to train SLT models. These methods usually rely on general datasets [47],

† Corresponding authors.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

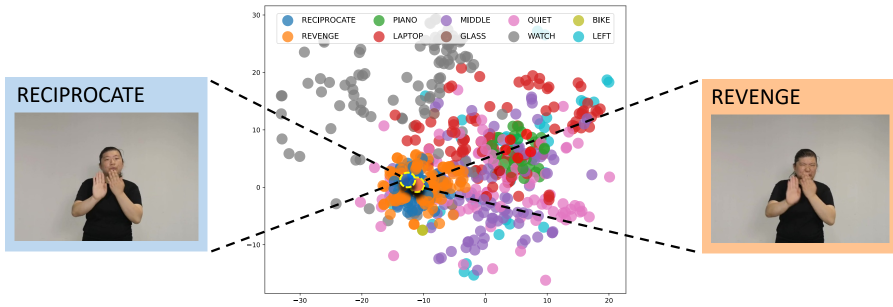


Figure 1: An example of the representation density problem in sign language translation. The two images show the sign gestures for "RECIPROCATE" (blue dot) and "REVENGE" (orange dot). Although the two have opposite meanings, their visual representations are densely clustered together, as shown in the t-SNE visualization. The various colors in the visualization indicate sign gestures with different meanings.

general pretraning strategy [52], or general large-scale foundation models [42] to promote gloss-free SLT. However, there is a substantial gap between the sign language domain and the general domain [45, 24]. Models trained with general strategies or datasets often fail to capture the subtle differences in semantically distinct gestures, which are crucial for accurately understanding a specific sign language. Therefore, the performance of gloss-free methods still significantly lags behind that of gloss-based approaches.

In this paper, we identify **a representation density problem** in sign language translation: the visual representations of sign gestures with distinct semantics are likely to be close in representation space. This problem is attributed to the nature of sign language, a form of visual language that utilizes intricate hand gestures, facial expressions, and body movements to convey the signer's message [35, 40, 17]. For example, in Figure 1, the signer performs sign gestures for opposite meanings, "RECIPROCATE" and "REVENGE", with similar visual information (i.e., only subtle differences in facial movements). The visual encoder in SLT models will encode similar visual information to visual representations in adjacent representation space, even though they have distinct semantics. Without explicit gloss annotations, SLT models struggle to learn semantic boundaries in continuous sign videos and capture distinguishing visual representations for different sign gestures. As a result, the representation density problem poses a significant challenge for the SLT models in distinguishing between various sign gestures, leading to sharp performance drops. (Section 3.2).

Further, we investigate various popular sign feature extraction methods, including gloss-based [3, 33] and gloss-free [47, 52], to systematically study the representation density problem. As shown in Figure 2, our investigation reveals that the representation density problem is prevalent across sign feature extraction methods. Specifically, due to the lack of gloss annotations, the representation density problem appears to be more serious in gloss-free methods. Then, we conduct extensive SLT experiments and observe that SLT models using gloss-free sign features as input consistently suffer a drop in performance in both sign language recognition and translation tasks compared to those using gloss-based sign features (Section 3.3). Therefore, we demonstrate that the representation density problem can be a bottleneck in restricting the improvement of gloss-free sign language translation.

More importantly, we propose a simple but effective contrastive learning strategy named SignCL to address the representation density problem. Specifically, SignCL draws the visual representations of sign gestures with identical semantics closer together and pushes those with different semantics farther apart. Experimental results show that SignCL can learn more distinctive feature representations and lead to significant improvements in terms of BLEU score on various well-known SLT frameworks (Section 5). To summarize, the main contributions of this work are as follows:

- To the best of our knowledge, our work identifies the representation density problem in sign language translation for the first time. This problem is consistent across various sign feature extraction methods for SLT, including gloss-based and gloss-free methods.
- Experimental results empirically reveal that an increase in representation density leads to a significant performance drop in the accuracy of sign language recognition and translation. We find that the representation density problem poses a significant challenge for the gloss-free SLT.

- We propose a simple but effective contrastive learning strategy, namely SignCL, to address the representation density problem. Our experiments demonstrate that SignCL can significantly enhance various well-known SLT frameworks. Specifically, SignCL yields a 39% BLEU score improvement for the Sign Language Transformer [4] and a 46% BLEU increase for GFSLT-VLP [53] on the CSL-Daily dataset.

2 Related Works

2.1 Sign Language Translation

Sign Language Translation (SLT) methods can be broadly categorized into gloss-based and gloss-free approaches. For gloss-based methods, an essential factor is to directly or indirectly employ sign gloss annotations to improve sign video encoder performance [3, 48, 53, 8, 49, 6]. These methods often employ Connectionist Temporal Classification [16] (CTC) loss to perform sign language recognition [4]. Joint-SLT [4] firstly introduces a multitask encoder-decoder framework with a CTCloss to soft-match sign representations and gloss sequences. STMC-T [54] introducing intra-cue and inter-cue CTC loss to model multi-cue sequence information. Despite their effectiveness, creating SLT datasets with gloss annotations is resource-intensive and time-consuming. Gloss-free methods have emerged as a promising alternative, as they do not rely on gloss annotations during training, making them more generalizable. And recently, a growing body of literature has promoted the gloss-free SLT, such as GASLT [47] proposed local gloss attention to mimic gloss assistant, GFSLT [52] adapted CLIP to do visual-language pretraining, and Sign2GPT [42] promoted performance by making use of large-scale pre-trained vision and language models. Nonetheless, the performance of gloss-free methods still significantly lags behind that of gloss-based approaches.

2.2 Contrastive Learning

Contrastive Learning [21, 55, 30], a popular unsupervised learning algorithm, aims to learn effective representations by pulling positive pairs closer together and pushing negative pairs farther apart. This approach has been widely utilized in both Natural Language Processing and Computer Vision [13]. In Sign Language Translation (SLT), Jin and Zhao [22] utilize Contrastive Learning to create a Signer-Independent SLT model, using videos demonstrating signs from different signers as positive samples. Additionally, Gan et al. [15] proposes a visual-level contrastive learning method with various image augmentation strategies. ConSLT [14] do contrastive learning for effective token representation learning in text decoder. Zhou et al. [52] and Cheng et al. [9] employ contrastive learning techniques to align video and text representations in SLT. In this paper, we are the first one to address the representation density problem, focusing particularly on visual gesture duration as a central aspect.

2.3 Representation Density

Representation Density is often a focal point in classification tasks, also known as category density [1, 43, 37, 31, 32, 51, 29, 12]. This concept pertains to the compactness and clarity of feature representations across different categories. In the context of sign language, various methods have been developed to address the subtle nuances of sign actions. TSPNet [28] proposes a temporal hierarchical attention network to learn segmented representations. HST-GNN [24] utilizes a hierarchical spatio-temporal graph neural network to learn graph representations from multiple perspectives. GLE-Net [20] employs global contextual relationships and fine-grained cues to distinguish non-manual-aware features in isolated Sign Language Recognition. These methods are beneficial for addressing the subtleties of sign language movements. However, integrating them into existing state-of-the-art frameworks presents significant challenges, often resulting in performance disparities when compared to the SOTA. This paper is the first to propose the concept of representation density within this field and introduces SignCL, which enhances the current mainstream transformer-based frameworks.

3 Representation Density Problem

This section investigates and identifies the representation density problem within existing sign feature extraction techniques, and examines whether representation density bottlenecks sign language recognition and translation performance.

3.1 Preliminaries

Existing Sign Feature Extraction Techniques Existing sign feature extraction methods can be divided into two categories: 1) gloss-based (e.g., Sign Recognition Pretrained [3] and Self-Mutual Knowledge Distillation [33]) and 2) gloss-free (e.g., I3D Pretraining [47] and Visual-Language Pretraining [52]). These methods were chosen for their representativeness in SLT and their well-documented open-source sign features.

- **Sign Recognition Pretrained (SRP)** [3]: This approach leverages the sign language recognition datasets to train sign language recognition models and uses it as the feature extractor for the SLT task. Notably, the features released by Camgoz et al. [3] have been widely adopted as input features in a range of works [5, 53, 22, 44, 46, 7].
- **Self-Mutual Knowledge Distillation (SMKD)** [18]: This approach enhances SRP by enforcing the visual and contextual modules to focus on short-term and long-term information [18]. SMKD feature extraction has been shown to substantially enhance SLT translation performance compared to SRP [50, 46].
- **I3D Pretraining (I3D)** [47]: This method employs I3D models as the backbone to pre-train the feature extractor, initially trained on the Kinetics dataset [25] and subsequently fine-tuned on extensive web SLR datasets, such as WSLR [27].
- **Visual-Language Pretraining (VLP)** [52]: This method entirely forgoes gloss annotations and leverages a general visual-language pretraining strategy to align sign video representation with text. Embodied by GFSLT-VLP [52], this approach offers a more general solution that utilizes a broader range of sign language resources without the constraints of gloss annotations.

Representation Density Metrics Drawing inspiration from Fisher's Discriminant Ratio (FDR) [23, 19], a typical measure used to evaluate the discriminative power of features in the classification, we combine the average Inter-Gloss Distance and Intra-Gloss Distance into Sign Density Ratio (SDR, see Eqn. 1), which reflects the degree of representation density for each gloss G_i. This is given by the formula:

$$SDR(G_i) = \frac{D_{G_i}^{intra}}{avg.D_{G_i}^{inter}} = \frac{D(G_i)}{Mean_{j \neq i}\left(D(G_i, G_j)\right)}. \tag{1}$$

Here, $D(G_i, G_j)$ represents the Inter-Gloss Distance between two glosses G_i and G_j, and avg. $avg.D_{G_i}^{inter}$ reflects the average distance of G_i to all other glosses. The Intra-Gloss Distance $D_{G_i}^{intra}$ evaluates the average distance within a single gloss G_i. These distances are given by the following formulas:

$$D(G_i, G_j) = \frac{1}{|G_i||G_j|} \sum_{x \in G_i, y \in G_j} d(x, y); \tag{2}$$

$$D(G_i) = \frac{1}{|G_i|(|G_i| - 1)} \sum_{x, y \in G_i, x \neq y} d(x, y); \tag{3}$$

Where, $|G_i|$ and $|G_j|$ denote the number of instances in glosses G_i and G_j respectively, and $d(x, y)$ represents the distance measure between the embeddings of instances x and y, i.e., euclidean distance.

The average Sign Density Ratio (SDR) of all glosses, denoted as $SDR = \text{Mean}(SDR(G_i))$, is calculated to evaluate the overall representation density of the dataset comprehensively.

Sign-Gloss Alignment To calculate the Sign Density Ratio (SDR), we need to determine the mapping relationship between input frames and gloss categories. Following previous works [26, 46], we employ the CTC classifier as a sign-gloss forced aligner to establish the mapping between each gloss and its corresponding sign frames. The aligner provides the start position l_v and end position r_v within the video frame sequence for each corresponding gloss g_v. To optimize alignment performance on the test set, we merge the training and test datasets for comprehensive training and engage two volunteers to select the best frame f_v from the range $[l_v{:}r_v]$ to align with each gloss g_v. Extensive details on the training procedure and the aligner's performance metrics are documented in Appendix 9.

3.2 Demonstrating Representation Density Problem

Experiment Setups We primarily use the PHOENIX-2014T benchmark [3] to investigate the representation density problem in existing sign feature extraction techniques. This benchmark was selected due to its rich collection of open-source sign features contributed by various research efforts. We obtained the sign features by either downloading the officially released versions or reproducing the feature extraction process. Then, we employed t-SNE [39] to visualize the feature distribution of these semantically distinct sign gestures to investigate representation density.

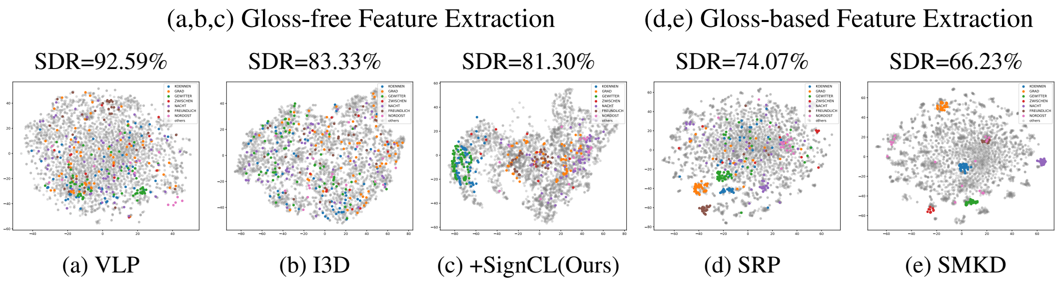


Figure 2: The t-SNE visualization of sign features across existing extraction techniques. SRP, SMKD, and I3D are downloaded from their official websites, while VLP is reproduced with official code. The addition of +SignCL denotes our proposed method that integrates a contrastive learning strategy into the VLP method (see Section 4). Different colors represent sign gestures with distinct semantics. Points in gray represent other sign categories not listed. Better viewed by zooming in.

Results and Findings Through empirical analysis of various visualized open-source sign features, we have identified a widespread representation density problem across different sign feature extraction methods. As depicted in Figure 2, all evaluated methods display a Sign Density Ratio exceeding 50%, with inevitable overlap of feature representation. Notably, gloss-free methods that do not utilize gloss annotations as additional supervision (e.g., I3D and VLP) exhibit even more severe representation density compared to gloss-based methods. This is evident as sign gestures representing different semantics, indicated by different colors, significantly overlap, resulting in translation ambiguity during inference. Specifically, the Sign Density Ratio (SDR) of VLP is 92.59%, which is significantly higher than the SDR of SMKD at 66.23%.

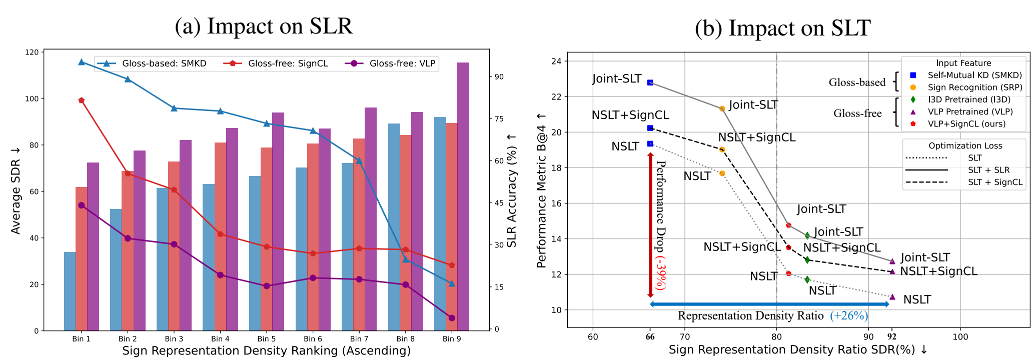


Figure 3: Comparative analysis of representation density and its impact on sign language recognition (SLR) and translation (SLT). The left panel (a) shows the correlation between representation density and SLR accuracy across different sign feature types and sign gesture groups. Binning in this context is based on sorting by gloss density within a group, where higher bins indicate higher density. The right panel (b) illustrates the performance drops in SLT caused by the representation density problem. This figure assesses both the recognition and translation accuracies, reflecting how denser representations impact these metrics.

3.3 Demonstrating Performance Drop

This section investigates the impact of representation density on sign language recognition (SLR) and translation (SLT) systems.

General Setups This part employs the widely utilized Sign Language Transformer [5] (NSLT) as the foundational model for our evaluations. The choice is because its capability to perform both SLT and SLR tasks, as well as take SLR and SLT at the same time (joint-SLT). Additionally, the NSLT framework is well-established within sign language research and benefits from comprehensive documentation and support in open-source sign feature sets and baseline results. The NSLT relies on sign features derived using a pretrained sign feature extractor. This section studies all types of sign features introduced in Section 3.2 to investigate the representation density problem. We use the Sign Density Ratio (SDR, see Eqn. 1) to measure the representation density within each type of input feature. We measure SLR and SLT performance by the recognition accuracy and the BLEU-4 [34] score (B@4), respectively.

Task Setups We set up tasks to examine whether representation density bottlenecks sign language recognition and translation performance.

- **Sign Language Recognition**: To evaluate the ability of the extracted sign features to distinguish between different semantic gestures, we use the NSLT [4] to perform sign language recognition (SLR) tasks with various types of sign features [3] as model input. Due to the limited number of samples for each gesture in the dev set, we rank the sign glosses based on their density using SDR(G_i) under SMKD features (see Eqn. 1). These glosses are then divided into nine groups (bins), each containing approximately 60 glosses. The average SDR(G_i) and recognition accuracy for each bin represents the overall density and mean accuracy of the glosses within that bin, respectively.
- **Sign Language Translation**: This evaluation aims to demonstrate the impact of representation density on translation tasks. We evaluate various sign features as inputs to the Sign Language Transformer, including SRP, SMKD, I3D, VLP, and VLP+SignCL. These inputs are tested across different translation frameworks, such as NSLT [3], Joint-SLT [4], and NSLT+SignCL. NSLT means use NSLT to perform SLT without CTC loss (gloss-free) and the NSLT+SignCL configuration integrates the proposed contrastive learning strategy into the encoder of NSLT [3] models, as detailed in Section 4.

Results and Findings As depicted in Figure 3, the following observations were made regarding the impact of representation density on both recognition (SLR) and translation (SLT):

- **Performance suffers from representation density.** We consistently observed a negative relationship between representation density and performance across all feature types and tasks. Higher representation density leads to worse accuracy in SLR and lower BLEU scores in SLT. Specifically, an increase in the representation density ratio by 26% can result in a 39% performance drop in NSLT.
- **Gloss-free methods suffer from worse representation density.** Gloss-free based feature extractions, which do not use any gloss annotations for assistance (e.g., VLP), typically exhibit higher representation density scores than gloss-based approach (e.g., SDR(VLP)=92.59% > SDR(SMKD)=66.23%). Using gloss-free features results in worse recognition and translation performance compared to gloss-based feature extractions (e.g., VLP vs. SMKD).
- **Contrastive learning boosts performance by reducing representation density.** When contrastive learning is applied to augment gloss-free based feature representation learning, i.e., VLP+SignCL for feature extraction or NSLT+SignCL for downstream finetuning, there is a consistent reduction in feature representation density accompanied by a significant improvement in both of the SLR accuracy and the SLT performance (see detail can be found in Section 4).

4 Contrastive Learning for Gloss-free Sign Langauge Translation

Contrastive Learning [21], a popular self-supervised learning algorithm, aims to learn effective representations by pulling positive pairs closer together and pushing negative pairs farther apart. In this section, we introduce a simple but efficient sign contrastive learning strategy, namelySignCL, which addresses the challenge of the representation density problem in gloss-free sign language translation.

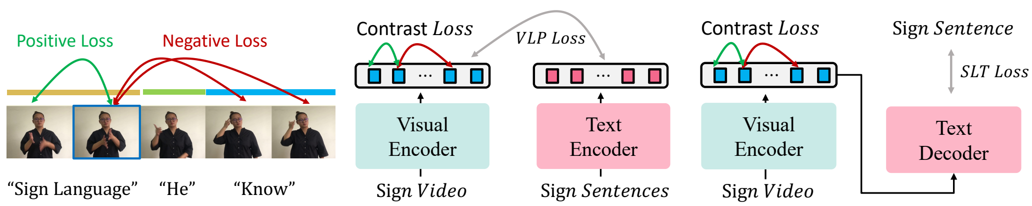
4.1 Sign Contrastive Learning

The key factor in contrastive learning is how to sample positive and negative training pairs. As illustrated in the framework shown in Figure 4a, the sampling strategy of SignCL is as follows: if two frames are close enough (e.g., adjacent), they are considered to belong to the same sign gesture and are treated as positive samples. Conversely, if two frames are far apart by double the margin (e.g., $|f_{ed} - f_{st}| > 20$ frames), they are considered to be associated with different semantics and are treated as negative samples. Statistically, the average duration of each gesture in sign video is nine frames [3, 53], and according to the speech-to-gesture Zipf's Law [2], each gloss represents approximately 2.3 spoken words. Therefore, we set the margin as $\max(10, \frac{\text{len(frames)}}{\text{len(text)}} \times 2.3)$.

$$\begin{cases} \text{positive pair } (f_{st}, f_{ed}^{+}): & |f_{ed}^{+} - f_{st}| \leq 1 \\ \text{negative pair } (f_{st}, f_{ed}^{-}): & |f_{ed}^{-} - f_{st}| > 2 * margin \end{cases}; \tag{4}$$

$$\mathcal{L}_{SignCL} = \frac{1}{N} \sum_{st=1}^{N} \left[\mathrm{d}(f_{st}, f_{ed}^{+}) + \max(0, m - \mathrm{d}(f_{st}, f_{ed}^{-})) \right]; \tag{5}$$

Where d is the distance function, i.e., Euclidean distance for frame features (f_{st}, f_{ed}), and N is the total number of frames in one sign video, $N = \text{len(frames)}$. The margin parameter m is used to prevent the features of the negative pair from being too far away. We empirically set $m = 64$ based on the average Inter-Gloss Distance (see Eqn. 2) of gloss-based sign features (e.g., SMKD[18]).



(a) Sign Contrastive Learning (b) SignCL in Pretraining (c) SignCL in Finetuning

Figure 4: Overview of the SignCL in gloss-free sign language translation: (a) Sign contrastive learning sampling strategy, (b) Showcases the integration of SignCL in the pretraining stage, and (c)) Displays the application of SignCL during the finetuning stage.

4.2 Integrating Contrastive Learning into Sign Language Translation Tranining

As illustrated in Figures 4b and 4c, *SignCL* can be integrated into both the sign feature extraction pretraining stage (e.g., Visual-Language Pretraining [52]) and the downstream task finetuning stage (e.g., GFSLT-VLP [52]). The optimization objective for these approaches is the weighted sum of $\mathcal{L}_{\text{SignCL}}$ and the original objective loss (e.g., VLP Loss for pretraining and SLT loss for finetuning [3, 52]), defined as:

$$\mathcal{L} = \lambda * \mathcal{L}_{\text{SignCL}} + \mathcal{L}_{MLE}; \tag{6}$$

Where L_{MLE} is the original objective loss in the pertaining or finetuning.

5 Experiments

In this Section, we conduct experiments to demonstrate the efficiency of proposed SignCL in reducing representation density and boosting gloss-free sign language translation performance. Specifically, we apply SignCL to the Sign Language Transformer [4] to facilitate a direct comparison with prior empirical analyses of the representation density problem in Section 3.3. Additionally, we integrate SignCL into the GFSLT-VLP [52] framework, a robust new gloss-free baseline that improves SLT through pretraining and finetuning.

5.1 Experiments on Sign Language Transformer

In Section 3, we investigate the representation density problem using the Sign Language Transformer (SLT) [4] and the PHOENIX-2014T [3] and CSL-Daliy [53] Dataset. These benchmarks are chosen for their established relevance in sign language translation research, including gloss-based and gloss-free based. Here, we first conduct experiments on the same framework and dataset to facilitate direct comparison with the prior empirical analyses.

Experiment Settings: In this experiment, we introduce SignCL as additional supervision information in the encoder of SLT under gloss-free settings. This enhanced model is referred to as +SignCL.

Results and Findings: The integration of SignCL into the SLT has significantly improved translation performance across all test conditions by reducing the representation density, as shown in Table 1. Notably, SignCL encourages SLT to learn a more distinct feature distribution, reducing the Sign Density Ratio (SDR) significantly, e.g., 66.23 to 62.18 and 92.59 to 81.30.

Figures 3a and 3b show experiments on SLR and SLT tasks using features with varying SDRs as inputs to SLT. The representation density reduction leads to observable improvements in both recognition accuracy (red line vs. purple line in Figure 3a) and translation BLEU score (purple point vs. red point in Figure 3b). Further details and additional experiment results on the CSL-Daily dataset are provided in Appendix A.4.3.

Table 1 presents a comparative analysis of representation density and performance on the PHOENIX-2014T dataset. The inclusion of SignCL during VLP feature extraction or SLT training processes significantly enhances performance metrics. WERs (Word Error Rates) in the gloss-free set, derived from an independent SLR task, are specifically used to probe the quality of sign features and do not participate in the SLT training process. This analysis underscores the significant enhancements brought by SignCL in terms of both efficiency and effectiveness in SLT frameworks.

Feature Type	**PHOENIX-2014T**			**CSL-Daily**		
	$SDR\downarrow$	WER↓	B@4↑	$SDR\downarrow$	WER↓	B@4↑
Gloss-based						
Joint-SLT [5] / Self-Mutual KD [33]	66.23	25.38	22.79	48.34	29.52	11.61
+ SignCL into Feature Extraction	62.18	24.76	23.23	-	-	-
+ SignCL into Finetuning	66.23	25.12	22.92	-	-	-
+ SignCL into both	**62.18**	**24.58**	**23.46**	-	-	-
Gloss-free						
SLT [3] / VLP Pretrained [52]	92.59	*69.72*	10.73	76.55	*85.78*	1.82
+ SignCL into Feature Extraction	81.30	*63.33*	12.04	68.39	*80.71*	2.15
+ SignCL into Finetuning	92.59	-	12.14	76.55	-	2.19
+ SignCL into both	**81.30**	-	**13.51**	**68.39**	-	**2.53**

Table 1: Comparative analysis of representation density and performance on the PHOENIX-2014T dataset. "+SignCL" indicates the inclusion of the proposed contrastive learning strategy during VLP (Video Language Processing) feature extraction or SLT (Sign Language Translation) training processes. WERs (Word Error Rates) in the gloss-free set are derived from an independent SLR (Sign Language Recognition) task, used specifically for probing the quality of sign features. These WERs do not participate in the SLT training process.

5.2 Experiments on Gloss-free Sign Language Translation

Gloss-free sign language translation, which does not rely on gloss annotations, has become a trend as it makes the approach more generalizable. In the realm of gloss-free sign language translation, GFSLT-VLP [52] stands out as a strong new baseline. It incorporates CLIP [36] and MBART [10] for model pretraining and finetuning. In this set of experiments, we use GFSLT-VLP as the baseline model and integrate the proposed SignCL into the framework to demonstrate the effectiveness of our method in both pretraining and finetuning settings.

Experiment Settings: This set of experiments is conducted using the PHOENIX-2014T [3] and CSL-Daily [53] datasets. We reproduce GFSLT-VLP using the official code and integrate SignCL into both the pretraining and finetuning stages. All models and training details are consistent with

Model	Density $SDR \downarrow$	Performance R@L↑	B@1↑	B@2↑	B@3↑	B@4↑
NSLT [3, 4]	-	30.07	29.86	17.52	11.96	9.00
GASLT [47]	-	39.86	39.07	26.74	21.86	15.74
GFSLT [52]	-	40.93	41.39	31.00	24.20	19.66
GFSLT-VLP [52]	-	42.49	43.71	33.18	26.11	21.44
Sign2GPT(w/PGP) [42]	-	48.90	49.54	35.96	28.83	22.52
GFSLT-VLP [52]	68.53	42.97	42.13	32.04	25.62	21.25
+ SignCL into Pretraining	62.68	**49.25**	**49.99**	36.73	29.76	22.69
+ SignCL into Finetuning	62.73	48.17	48.56	35.04	27.73	22.16
+ SignCL into Two State	**62.32**	49.04	49.76	**36.85**	**29.97**	**22.74**
Improvement	-6.21	+6.07	+7.63	+4.81	+4.35	+1.49

Table 2: Improvement in the GFSLT-VLP framework by reducing representation density on PHOENIX-2014T test set. "+SignCL into Pretraining" indicates applying the proposed contrastive learning strategy during the pretraining stage, while "+SignCL into Finetuning" indicates the inclusion of the SignCL during the finetuning stage. "+SignCL into Two State" means plus SignCL both in pertaining and finetuning states.

Model	Density $SDR \downarrow$	Performance R@L↑	B@1↑	B@2↑	B@3↑	B@4↑
GASLT [47]	-	20.35	19.90	9.94	5.98	4.07
NSLT [3, 4]	-	34.54	34.16	19.57	7.56	7.56
GFSLT [52]	-	35.16	37.69	23.28	14.93	9.88
GFSLT-VLP [52]	-	36.44	39.37	24.93	16.26	11.00
Sign2GPT(w/PGP) [42]	-	42.36	41.75	28.73	20.60	15.40
GFSLT-VLP [52]	58.20	39.08	36.37	23.32	15.45	11.10
+ SignCL into Pretraining	55.24	47.38	46.20	32.33	22.35	15.85
+ SignCL into Finetuning	55.03	48.26	46.53	32.41	22.42	15.98
+ SignCL into Two States	**54.61**	**48.92**	**47.47**	**32.53**	**22.62**	**16.16**
Improvement	-3.59	+9.84	+11.10	+9.21	+7.17	+5.06

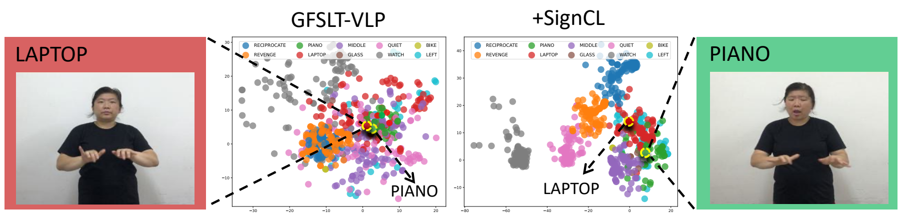
Table 3: Enhancing GFSLT-VLP by reducing representation density on CSL-Daily test set.

those used in GFSLT-VLP [52], with the sole exception being the incorporation of SignCL, weighted by $\lambda = 0.01$, as illustrated in Figure 2 and Equation 6. Further details are provided in Appendix A.1.

Results and Findings: Tables 2 and 3 compare our proposed methods with existing gloss-free sign language translation approaches. The results demonstrate that integrating the proposed SignCL strategy into the GFSLT-VLP framework consistently reduces representation density and significantly boosts translation performance, whether SignCL is applied during pretraining, finetuning, or both stages. Specifically, compared to the baseline model GFSLT-VLP [52], our approach achieves a substantial improvement of 45.58% (+5.06) in the BLEU-4 score on the CSL-Daily dataset, without any increase in the number of parameters. Additionally, despite having significantly fewer parameters ($\sim$600M vs. $\sim$1.7B), our approach achieves better performance than Sign2GPT [42], which leverages large-scale pretrained vision and language models for sign language translation.

5.3 Qualitative Analysis

To understand our SignCL approach in scenarios of addressing representation density, we present a case from the CSL-Daily dataset in Figure 5. As shown, the way to display sign gestures for "电脑" (laptop) and "钢琴" (piano) differ subtly. As indicated by the t-SNE results, the representations of these two semantically different gestures are closely packed together in the feature space, causing the baseline GFSLT-VLP model to incorrectly translate "钢琴" (piano) as "电脑" (laptop). In contrast, our proposed SignCL effectively separates the representations of "电脑" (laptop) and "钢琴" (piano) in the feature space, enabling the accurate translation of "钢琴" (piano).



Reference	妈妈给我买钢琴。(Mom bought me a piano.)
GFSLT-VLP	妈妈给我买笔记本电脑。(Mom bought me a laptop.)
GFSLT-VLP + SignCL	妈妈给我买钢琴。(Mom bought me a piano.)

Figure 5: Qualitative comparison of translation results on CSL-Daily test set. The red background denotes model misinterpretations about the sign gestures, while green one means accurate recognition. Content in (...) is English translation for non-Chinese readers.

6 Conclusion

In this work, we identify a crucial representation density problem in gloss-free sign language translation. Our systematic investigation reveals that this problem persists across various existing sign feature extraction methods and causes sharp performance drops in both sign language recognition and translation, particularly in gloss-free methods. To address this problem, we propose a simple but effective contrastive learning strategy, termed SignCL. Our experiments demonstrate that SignCL encourages gloss-free models to learn more discriminative features and significantly reduces representation density. Furthermore, our experiments show that SignCL improves translation performance across various frameworks and datasets by a significant margin, achieving a new state-of-the-art in gloss-free sign language translation. We illustrate the effectiveness of SignCL through detailed examples in our qualitative analysis. Finally, we provide several ablation studies for a better understanding of SignCL and discuss the limitations and potential societal impacts of this work in the Appendix A.

7 Limitations

Our work, while promising, has several limitations that should be considered:

Boundary Cases: We assume that adjacent frames output the same sign gestures, while distant frames belong to different sign gestures. This assumption might not hold in special sign language videos with extensive repetitive gestures. In extreme cases, SignCL might affect feature convergence.

Semantic Similarity: SignCL does not account for the semantic similarity between sign gestures, which can result in increased feature distances between semantically similar gestures. This could potentially affect the learning of linguistic features.

Despite acknowledging these limitations, our experiments demonstrate that our approach works effectively in most cases. We will address these issues in future work to further enhance the robustness and applicability of our method.

8 Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (Grant No. 92370204), in part by the National Key R&D Program of China (Grant No. 2023YFF0725001), in part by the Guangzhou-HKUST (GZ) Joint Funding Program (Grant No. 2023A03J0008), and in part by the Education Bureau of Guangzhou Municipality. This work was also supported by the Guangzhou Municipal Science and Technology Project (No. 2024A04J4390).

References

[1] Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

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

[3] Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018). Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793.

[4] Camgoz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2020a). Multi-channel transformers for multi-articulatory sign language translation. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 301–319. Springer.

[5] Camgoz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2020b). 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.

[6] Chen, C., Guo, W., Ma, C., Yang, Y., Wang, Z., and Lin, C. (2021a). semg-based continuous estimation of finger kinematics via large-scale temporal convolutional network. *Applied Sciences*, 11(10):4678.

[7] Chen, R., Chen, Y., Guo, W., Chen, C., Wang, Z., and Yang, Y. (2021b). Semg-based gesture recognition using gru with strong robustness against forearm posture. In *2021 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, pages 275–280. IEEE.

[8] Chen, Y., Wei, F., Sun, X., Wu, Z., and Lin, S. (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.

[9] Cheng, Y., Wei, F., Bao, J., Chen, D., and Zhang, W. (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.

[10] Chipman, H. A., George, E. I., McCulloch, R. E., and Shively, T. S. (2022). mbart: multidimensional monotone bart. *Bayesian Analysis*, 17(2):515–544.

[11] Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

[12] Dai, L., Liu, H., and Xiong, H. (2024). Improve dense passage retrieval with entailment tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

[13] Dai, L., Ma, L., Qian, S., Liu, H., Liu, Z., and Xiong, H. (2023). Cloth2body: Generating 3d human body mesh from 2d clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15007–15017.

[14] Fu, B., Ye, P., Zhang, L., Yu, P., Hu, C., Shi, X., and Chen, Y. (2023). A token-level contrastive framework for sign language translation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[15] Gan, S., Yin, Y., Jiang, Z., Xia, K., Xie, L., and Lu, S. (2023). Contrastive learning for sign language recognition and translation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 763–772.

[16] Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 369–376.

[17] Guo, W., Li, Z., Yang, Y., Wang, Z., Taylor, R. H., Unberath, M., Yuille, A., and Li, Y. (2022). Context-enhanced stereo transformer. In *European Conference on Computer Vision*, pages 263–279. Springer.

[18] Hao, A., Min, Y., and Chen, X. (2021). Self-mutual distillation learning for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11303–11312.

[19] Harish, B. and Manju, N. (2018). Hybrid feature selection method using fisher's discriminate ratio to classify internet traffic data. In *Proceedings of the 4th International Conference on Frontiers of Educational Technologies*, pages 75–79.

[20] Hu, H., Zhou, W., Pu, J., and Li, H. (2021). Global-local enhancement network for nmf-aware sign language recognition. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 17(3):1–19.

[21] Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F. (2020). A survey on contrastive self-supervised learning. *Technologies*, 9(1):2.

[22] Jin, T. and Zhao, Z. (2021). Contrastive disentangled meta-learning for signer-independent sign language translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5065–5073.

[23] Kalinkov, K., Ganchev, T., and Markova, V. (2019). Adaptive feature selection through fisher discriminant ratio. In *2019 International Conference on Biomedical Innovations and Applications (BIA)*, pages 1–4.

[24] Kan, J., Hu, K., Hagenbuchner, M., Tsoi, A. C., Bennamoun, M., and Wang, Z. (2022). Sign language translation with hierarchical spatio-temporal graph neural network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3367–3376.

[25] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

[26] Kürzinger, L., Winkelbauer, D., Li, L., Watzel, T., and Rigoll, G. (2020). Ctc-segmentation of large corpora for german end-to-end speech recognition. In *International Conference on Speech and Computer*, pages 267–278. Springer.

[27] Li, D., Rodriguez, C., Yu, X., and Li, H. (2020a). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469.

[28] Li, D., Xu, C., Yu, X., Zhang, K., Swift, B., Suominen, H., and Li, H. (2020b). Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *Advances in Neural Information Processing Systems*, 33:12034–12045.

[29] Lin, C., Chen, X., Guo, W., Jiang, N., Farina, D., and Su, J. (2022). A bert based method for continuous estimation of cross-subject hand kinematics from surface electromyographic signals. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:87–96.

[30] Lin, K.-Y., Ding, H., Zhou, J., Peng, Y.-X., Zhao, Z., Loy, C. C., and Zheng, W.-S. (2024). Rethinking clip-based video learners in cross-domain open-vocabulary action recognition. *arXiv preprint arXiv:2403.01560*.

[31] Liu, W., Wen, Y., Yu, Z., and Yang, M. (2016). Large-margin softmax loss for convolutional neural networks. *arXiv preprint arXiv:1612.02295*.

[32] Ma, C., Guo, W., Zhang, H., Samuel, O. W., Ji, X., Xu, L., and Li, G. (2021). A novel and efficient feature extraction method for deep learning based continuous estimation. *IEEE Robotics and Automation Letters*, 6(4):7341–7348.

[33] Min, Y., Hao, A., Chai, X., and Chen, X. (2021). Visual alignment constraint for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11542–11551.

[34] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

[35] Pizzuto, E. A. (2003). Review of "the hands are the head of the mouth—the mouth as articulator in sign languages" by penny boyes braem and rachel sutton-spence (eds.). *Sign Language & Linguistics*, 6(2):284–289.

[36] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

[37] Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

[38] Sedgwick, P. (2014). Spearman's rank correlation coefficient. *Bmj*, 349.

[39] Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11).

[40] Vinson, D. P., Thompson, R. L., Skinner, R., Fox, N., and Vigliocco, G. (2010). The hands and mouth do not always slip together in british sign language: Dissociating articulatory channels in the lexicon. *Psychological Science*, 21(8):1158–1167.

[41] Wang, Y., Liu, S., Wang, M., Liang, S., and Yin, N. (2024). Degree distribution based spiking graph networks for domain adaptation. *arXiv preprint arXiv:2410.06883*.

[42] Wong, R., Camgoz, N. C., and Bowden, R. (2024). Sign2GPT: Leveraging large language models for gloss-free sign language translation. In *The Twelfth International Conference on Learning Representations*.

[43] Xie, J., Girshick, R., and Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR.

[44] Yao, H., Zhou, W., Feng, H., Hu, H., Zhou, H., and Li, H. (2023). Sign language translation with iterative prototype. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15592–15601.

[45] Ye, J., Jiao, W., Wang, X., and Tu, Z. (2023a). Scaling back-translation with domain text generation for sign language gloss translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 463–476, Dubrovnik, Croatia. Association for Computational Linguistics.

[46] Ye, J., Jiao, W., Wang, X., Tu, Z., and Xiong, H. (2023b). Cross-modality data augmentation for end-to-end sign language translation. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13558–13571, Singapore. Association for Computational Linguistics.

[47] Yin, A., Zhong, T., Tang, L., Jin, W., Jin, T., and Zhao, Z. (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.

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

[49] Zhang, B., Müller, M., and Sennrich, R. (2023a). Sltunet: A simple unified model for sign language translation. In *International Conference on Learning Representations*.

[50] Zhang, B., Müller, M., and Sennrich, R. (2023b). Sltunet: A simple unified model for sign language translation. *arXiv preprint arXiv:2305.01778*.

[51] Zhang, H., Sun, Y., Guo, W., Liu, Y., Lu, H., Lin, X., and Xiong, H. (2023c). Interactive interior design recommendation via coarse-to-fine multimodal reinforcement learning. *arXiv preprint arXiv:2310.07287*.

[52] Zhou, B., Chen, Z., Clapés, A., Wan, J., Liang, Y., Escalera, S., Lei, Z., and Zhang, D. (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.

[53] Zhou, H., Zhou, W., Qi, W., Pu, J., and Li, H. (2021a). 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.

[54] Zhou, H., Zhou, W., Zhou, Y., and Li, H. (2021b). Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia*, 24:768–779.

[55] Zhou, J., Ma, T., Lin, K.-Y., Qiu, R., Wang, Z., and Liang, J. (2024). Mitigating the human-robot domain discrepancy in visual pre-training for robotic manipulation. *arXiv preprint arXiv:2406.14235*.

A Appendix

A.1 Hyper-parameters of Baselines

Sign Language Transformers Baseline: Table 4 presents the hyper-parameters of Sign Language Transformers used in this work.

Parameter	PHOENIX-2014T	CSL-Daily
encoder-layers	3	1
decoder-layers	3	1
attention heads	8	8
ctc-layers	1	1
hidden size	512	512
activation function	gelu	gelu
learning rate	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$
Adam β	(0.9, 0.98)	(0.9, 0.98)
label-smoothing	0.1	0.1
max output length	30	50
dropout	0.3	0.3
batch-size	128	128

Table 4: Hyperparameters of Sign Language Transformer models.

Gloss-Free Sign Language Translation Baseline: The Gloss-Free Sign Language Translation (GFSLT) model incorporates various modules designed for processing sign language input without the use of glosses. Below is the detailed architecture used in this work:

Module	Stride	Kernel	Output Size
Sign Input	-	-	$B \times T \times 224 \times 224 \times 3$
Resnet w/o fc	-	-	$B \times T \times 512$
Conv1D-BN1D-RELU	1	5	$B \times T \times 1024$
MaxPooling1D	2	2	$B \times \frac{T}{2} \times 1024$
Conv1D-BN1D-RELU	1	5	$B \times \frac{T}{2} \times 1024$
MaxPooling1D	2	2	$B \times \frac{T}{4} \times 1024$
Linear-BN1D-RELU	-	-	$B \times \frac{T}{4} \times 1024$
Transformer Encoder	-	-	$B \times \frac{T}{4} \times U$
Text Input	-	-	$B \times U$
Word Embedding	-	-	$B \times U \times 1024$
Transformer Decoder	-	-	$B \times U \times 1024$
FC	-	-	$B \times U \times C$

Table 5: Detailed Gloss-Free SLT (GFSLT) Framework. B represents batch size, T denotes the length of the longest input sign video in the batch, and U is the length of the longest input text in the batch. It is copied from GFSLT-VLP [52].

A.2 Parameter Sensitivity Analysis of the SignCL

A.2.1 Sensitivity Analysis on Dynamically Estimated Margin

The margin for negative sampling dynamically depends on the estimated average margin of each gloss, calculated as len(frames)$/$len(text) $\times$ speech-to-gesture Zipf's factor, with a minimum threshold set at 10. The Zipf's factor, set at 2.3, refers to the speech-to-gesture application of Zipf's Law.

We calculated the distribution of the dynamically estimated margin, with the results displayed in the table below. A more detailed distribution can be seen in Figure 6.

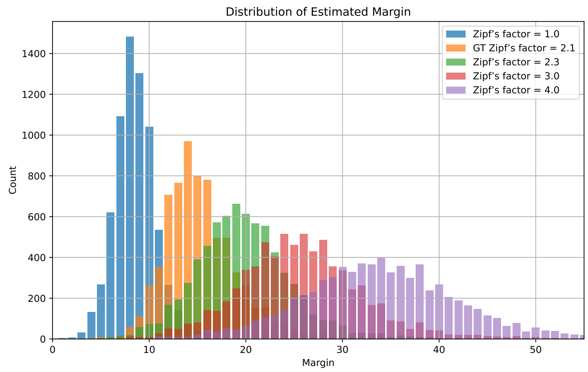


Figure 6: The distribution of the estimated margin during training on the PHOENIX-2014T dataset. The green distribution represents our current paper's method (factor = 2.3), while the orange distribution shows the ground truth calculated based on gloss annotations.

Experiment Setup: To conduct a principled analysis, we evaluated the threshold values at $[0, 10, 20, 30, 40, 50]$. Here, a threshold of 0 indicates that the margin is dominated by the dynamically estimated margin, while a threshold of 50 suggests dominance by the fixed threshold.

Experiment Results: We uniformly trained for 80 epochs on the PHOENIX-2014T dataset due to resource limitations. The results, as shown in the table below, indicate that SignCL is not sensitive to the threshold parameter, with a variance of 0.062.

Threshold	0	10	20	30	40	50
B@4	17.24	17.63	17.55	17.63	17.13	17.11

Table 6: Threshold sensitivity analysis results.

Zipf's factor	1	GT	2.3	3	4
B@4	17.45	17.89	17.63	17.29	16.26

Table 7: Sensitivity to Zipf's factor.

A.2.2 Sensitivity Analysis on integrating SignCL into the SLT framework.

As shown in Eqn. 6, we vary the hyperparameter λ over the range $[10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}]$ and conduct repeated experiments on the PHOENIX-2014T dataset with GFSLT-VLP.

As shown in Figure 7, excessively incorporating SignCL into the model can negatively impact the SLT task. Empirically, we find that $\lambda = 10^{-2}$ achieves a balance between reducing representation density and improving translation performance.

A.3 Ablation Studies

We conduct ablation studies to investigate the impact of different loss components in the +SignCL approach during both the pretraining and fine-tuning stages. It is copied from Tabel 2 , but with an ablation perspective.

A.4 Correlation between Representation Density and Recognition Performance

A.4.1 The efficiency of sign-gloss mapping building up

To calculate SDR, we need to establish the mapping relationship between input frames and gloss categories. This section presents the performance of our trained gloss-sign aligner. The experimental methodology follows the approach outlined in XmDA [46].

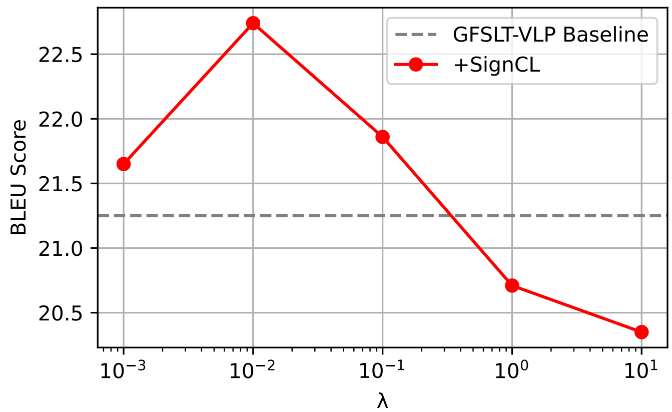


Figure 7: The effect of the hyperparameter λ on BLEU scores. the grey dashed line indicates the baseline performance of GFSLT-VLP, while the red solid line represents the performance with SignCL integrated.

Pretraining Stage		**Finetuning Stage**		**Density**	**Performance**	
VLP Loss	**SignCL Loss**	**SLT Loss**	**SignCL Loss**	$SDR\downarrow$	R@L↑	B@4↑
✗	✗	✓	✗	72.83	38.67	18.53
✗	✓	✓	✗	63.23	39.12	18.71
✓	✗	✓	✗	68.53	42.97	21.25
✓	✓	✓	✗	62.68	49.25	22.69
✗	✗	✓	✓	69.54	41.78	19.81
✗	✓	✓	✓	63.67	44.52	20.03
✓	✗	✓	✓	62.73	48.17	22.16
✓	✓	✓	✓	**62.32**	**49.23**	**22.74**

Table 8: Ablation study on the impact of different loss components in the +SignCL approach.

A.4.2 Correlation Coefficient

We analyze the relationship between the representation density of individual glosses and their recognition accuracy using the Sign Language Transformer on the PHOENIX-2014T dataset, leveraging the Self-Mutual Knowledge Distillation (SMKD) method for feature extraction. We compute the following correlation coefficients:

Pearson Correlation Coefficient [11]: This measures the linear relationship between recognition accuracy $Acc(G_i)$ for gloss G_i and the density metric $SDR(G_i)$, calculated as:

$$r = \frac{\sum(Acc(G_i) - \bar{Acc})(SDR(G_i) - S\bar{D}R)}{\sqrt{\sum(Acc(G_i) - \bar{Acc})^2 \sum(SDR(G_i) - S\bar{D}R)^2}} \tag{7}$$

Spearman's Rank Correlation Coefficient [38]: This assesses the monotonic relationship between two datasets by considering the rank order of values.

Dataset	**WER↓**		
	Train	Test	Dev
PHOENIX-2014T	8.68	8.03	25.28
CSL-Daily	9.23	8.39	29.32

Table 9: Evaluation of the gloss-sign aligner effectiveness and generalizability with WER (%) (the lower the better).

As shown in Table 10, both correlation coefficients indicate a medium negative correlation between the Sign Density Ratio (SDR) and sign recognition performance (Acc). This suggests that higher representation density correlates with poorer recognition performance. Specifically, Inter-Gloss Distance ($D_{G_i}^{inter}$) shows a strong positive correlation, meaning that greater distances between different glosses correlate with better recognition performance. All Spearman P-values are lower than 0.01, confirming the high confidence in the non-randomness of these correlations.

Correlation / Acc	**PHOENIX-2014T**			**CSL-Daily**		
	$D_{G_i}^{inter} \uparrow$	$D_{G_i}^{intra} \downarrow$	SDR $\downarrow$	$D_{G_i}^{inter} \uparrow$	$D_{G_i}^{intra} \downarrow$	SDR $\downarrow$
Pearson r	0.36	-0.22	-0.35	0.30	-0.14	-0.20
Spearman ρ	0.43	-0.24	-0.34	0.45	-0.16	-0.22
P-value	2.6E-17	5.5E-6	4.7E-11	6.6E-19	3.0E-3	-2.7E-5

Table 10: Correlation analysis between sign recognition performance and density metrics.

A.4.3 More Experiment Results on Sign Language Transformer

In this section, we present additional experimental results using the Sign Language Transformer (NSLT) on the CSL-Daily dataset to further validate the effectiveness of the proposed SignCL strategy. We compare various feature extraction methods to assess their representation density and translation performance.

Feature Type	**Density**	**Performance**			
	$SDR \downarrow$	SLR(WER $\downarrow$)	Joint-SLT	NSLT	+SignCL(ours)
Gloss-based Feature Extraction					
Sign Recognition [5]	74.07	29.59	21.32	17.68	19.02
Self-Mutual KD [33]	**66.23**	**25.38**	**22.79**	**19.35**	**20.23**
Gloss-free Feature Extraction					
I3D Pretrained [47]	83.33	61.74	14.17	11.70	12.81
VLP Pretrained [52]	92.59	69.72	12.73	10.73	12.14
+ SignCL (ours)	**81.30**	**63.33**	**14.76**	**12.04**	**13.51**

Table 11: Comparative analysis of representation density and performance on the PHOENIX-2014T dataset. "+SignCL (ours)" indicates the inclusion of the proposed contrastive learning strategy during VLP feature extraction or NSLT training processing.

Feature Type	**Density**	**Performance**			
	$SDR \downarrow$	SLR(WER $\downarrow$)	Joint-SLT	NSLT	+SignCL(ours)
Gloss-based Feature Extraction					
Self-Mutual KD [33]	**48.34**	**29.52**	**11.61**	**8.97**	**10.35**
Gloss-free Feature Extraction					
VLP Pretrained [52]	76.55	85.78	2.93	1.82	2.19
+ SignCL (ours)	**68.39**	**80.71**	**3.18**	**2.29**	**2.53**

Table 12: Comparative analysis of representation density and performance on the CSL-Daily dataset. The Self-Mutual KD features are provided by XmDA [46] and the VLP feature is reproduced with official code. Due to the incomplete open source of the CSL-Daily dataset, we were unable to obtain features for Sign Recognition and I3D Pretraining.

A.5 Broader Impacts

This paper focuses on research in sign language translation, which has the potential to significantly benefit individuals who are deaf or hard of hearing. By improving the accuracy and efficiency of sign language translation, our work can facilitate better communication between individuals with hearing impairments and the broader community. This can help break down communication barriers, promoting inclusivity and equal opportunities in various social, educational, and professional settings.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the main contributions and scope of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the Appendix, we discuss the limitations of this work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We disclose all the information needed to reproduce the main experimental results in the Methodology, Experiment, and Appendix sections. We will release the code upon acceptance of the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We commit to releasing the code and models upon acceptance of the paper. All the data and baselines are based on open-source benchmarks.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all the training and test details necessary to understand the results in the Methodology, Experiment, and Appendix sections.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported because performing multiple runs for each experiment would be too computationally expensive.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments are conducted using PyTorch on 8*NVIDIA A800 GPUs for about 12 hours.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer:[Yes]

   Justification: We adhere to the NeurIPS Code of Ethics, since the paper does not include any content or practices that violate ethical guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential societal impacts of the paper in the Appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not involve data or models that have a high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly credits the creators or original owners of all used assets, and properly respects the license and terms of use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets are introduced in the paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects, thus there are no study participants, no risks to disclose to subjects, and no need for Institutional Review Board (IRB) approvals.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.