Learning Effective Sign Features without Text for Gloss-free Sign Language Translation

Shiwei Gan[1†] Xiao Liu[1†] Yafeng Yin[1*] Nan Liu[1] Kuizhuang Liu[1] Desibieer Tuerdaken[1]
Zhiwei Jiang[1] Lei Xie[1] Sanglu Lu[1] Hongkai Wen[2]

[1]State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing 210023, China

[2]Department of Computer Science, The University of Warwick, UK

sw@nju.edu.cn, liuxiaox@smail.nju.edu.cn, yafeng@nju.edu.cn

{nanliu,liukz,des}@smail.nju.edu.cn, {jzw,lxie,sanglu}@nju.edu.cn, hongkai.wen@warwick.ac.uk

Abstract

Self-supervised learning (SSL) has achieved remarkable success across both NLP and CV domains. However, sign language translation (SLT) models still heavily rely on gloss annotations in gloss-based SLT or text annotations in gloss-free SLT (GFSLT) during pretraining, aiming to ensure that the backbone provides effective sign language (SL) features for the translation model. Such reliance restricts the scalability and generalization ability of the SLT model. One natural question arises: ***Can existing SSL methods be directly applied to the SL domain to train an effective sign feature extractor for downstream GFSLT tasks, eliminating the need for text annotations?*** *In this paper, we propose a simple yet effective pretraining framework with two goals: (1) decoupling the pretraining process from gloss or text annotations, relying purely on sign frames; and (2) only global frames are required during inference for simplicity. We show that directly applying existing SSL methods yields suboptimal performance, as SL features involve subtle motion patterns and discriminative cues that are often confined to local regions. To achieve this, we introduce SignDINO, a simple yet effective sign-aware DINO training strategy that learns effective and semantically meaningful representations from global frames without any textual supervision. Specifically, a teacher-student architecture is employed, where the teacher model receives the global sign frame, while the student model learns from masked local views that preserve only the hand and facial regions. Such a simple design encourages the model to infer global semantics from discriminative local cues, allowing the teacher model to extract SL-related features during inference solely based on global views. Extensive experiments on public SL datasets show that SignDINO achieves highly competitive performance on the GFSLT task without relying on extra cues or additional SL-related datasets pretraining.*

1. Introduction

Current sign language translation (SLT) tasks [8, 36] mainly include: Gloss-based Sign Language Translation and Gloss-free Sign Language Translation (GFSLT). The Gloss-based SLT [4, 10, 45] follows such de-facto paradigms: pretraining the SL tokenizer[1] under Continuous Sign Language Recognition (CSLR) (*i.e.*, pretraining the SL tokenizer with gloss labels under CTC constraints), and then finetuning with a translation model (*e.g.*, mbart [12, 16], GPT-2 [44], mT5 [48]) to generate the corresponding spoken language sentence. Previous state-of-the-art gloss-based SLT models [4, 10] have emphasized that pretraining with gloss labels highly boosts SLT performance. However, due to the dependency on gloss labels, SLT has increasingly shifted its focus towards GFSLT [12, 25, 44], which aims to boost SLT performance in the absence of gloss supervision.

As shown in Figure 1, to reduce the dependence on gloss annotations in pretraining SL tokenizers for SLT tasks, text annotations have become a widely adopted alternative. Therefore, current GFSLT approaches heavily rely on text supervision to pretrain their SL tokenizers. Several methods have been proposed to address the alignment challenge between visual and textual representations, such as contrastive learning with text [44], contrastive learning with pseudo glosses (generated from text), CTC constraint with pseudo glosses [12], multi-stage pretraining with text [13, 36], and quantization-based approaches conditioned on text.

This raises a fundamental question: **Are text annotations truly necessary for pretraining SL tokenizers in gloss-free SLT?** In other words, can we pretrain the visual

[†]Equal contribution

[*]Yafeng Yin is the corresponding author.

[1]For convenience, we use the term SL tokenizer to refer to the visual encoder that extracts sign language video features.

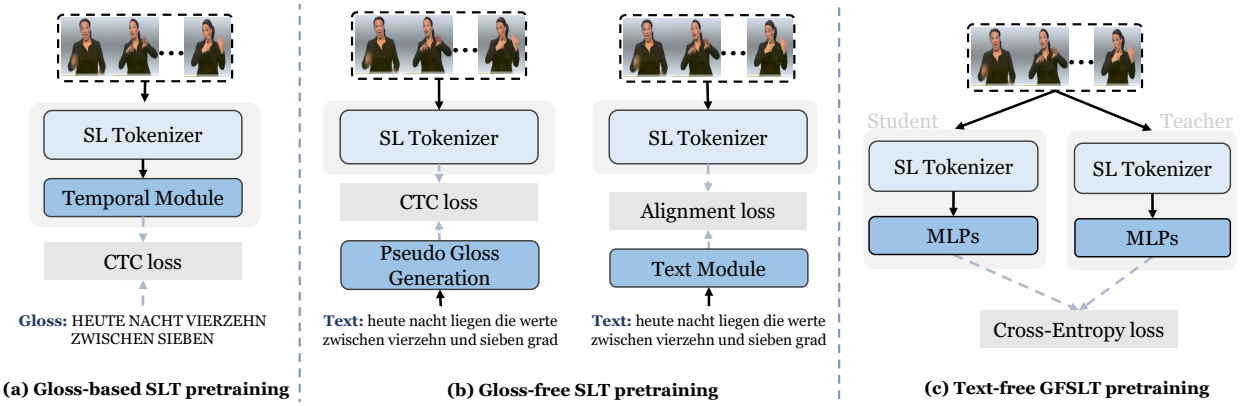


Figure 1. Comparison between gloss-based SLT, current gloss-free SLT and our text-free GFSLT pretraining. Note that other GFSLT models use text in both pretraining and finetuning. Our "Text-free" means no text annotations are used in backbone pretraining, while text annotations are still used for GFSLT finetuning, as in all other methods.

backbone solely from sign videos without any text supervision? **Our answer is that text supervision may not be necessary.** Although recent works such as SHuBERT [16] and SignMusketeers [15] have adopted DINO-like pretraining strategies by fine-tuning DINO encoders as their visual backbones without text, their approaches still exhibit two key limitations: (1) the pretraining focuses on localized discriminative regions (*e.g.*, hands and faces) rather than the full-frame representations, and (2) during inference, the visual backbones can only take these localized regions as input, instead of utilizing global video frames.

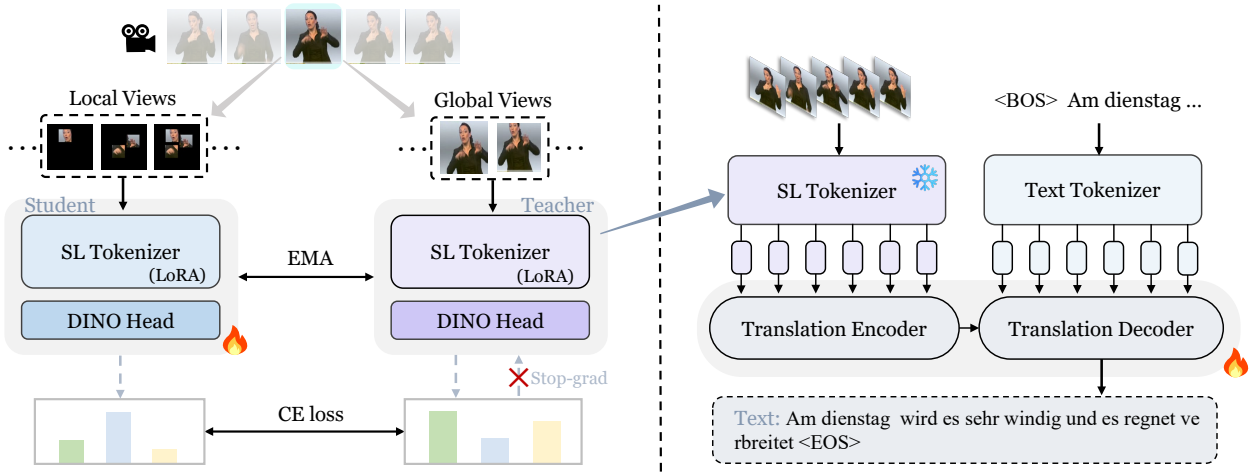
In this paper, we revisit existing SSL methods in the CV domain to examine whether they are applicable to GFSLT tasks. Furthermore, we aim to develop a general pretraining strategy for GFSLT with two main objectives: **(1) to decouple the pretraining process from gloss annotations and text annotations for improving scalability, allowing the backbone to utilize even raw, unlabeled sign videos**, and **(2) during inference, the SL tokenizer (*i.e.*, visual encoder) should only depend on global video frames without requiring extra inputs such as localized areas (hands, face) or skeletons.** To achieve these goals, we first investigate several state-of-the-art SSL methods, including generative unsupervised learning (*e.g.*, MAE) and contrastive unsupervised learning. We observe that directly applying these general-purpose training strategies to SL videos often leads to suboptimal results. This is mainly because such SSL algorithms tend to capture global semantic features rather than focusing on the fine-grained, discriminative local cues. In SL videos, however, most frames share similar global content (*e.g.*, background and body appearance), while the most informative and discriminative features are concentrated in the hand and facial regions.

To tackle the above challenges, we draw inspiration from the DINO series [32, 39] and design a simple, straightforward yet effective pretraining strategy combined with a sign-aware preprocessing scheme. Specifically, we introduce a sign-aware data augmentation strategy, where the teacher model receives the full global frame, while the student model learns from a masked local frame that only retains the hand and facial regions. This encourages the teacher model to infer discriminative local cues from global views. Our main contributions are summarized as follows:

- We investigate the effectiveness of current state-of-the-art SSL methods when directly applied to sign language pretraining, and reveal that existing approaches struggle to capture the discriminative local features essential for sign understanding.
- We propose a simple, straightforward and effective **sign-aware DINO** pretraining strategy that enables fully gloss-free and text-free sign language pretraining. The proposed method effectively learns discriminative sign representations from global video frames by self-distillation with different data views.
- We propose a simple GFSLT model: SignDINO, which utilizes only global frames to extract effective SL features. Extensive experiments on public SL datasets demonstrate its effectiveness, achieving excellent performance on the GFSLT task across multiple SL datasets.

2. Related Work

Self-supervised Learning. Self-supervised learning (SSL) aims to learn powerful data representations without the need for manual annotations, typically through so-called pretext tasks (or proxy tasks). In the NLP domain, pretext tasks such as next-token prediction (*e.g.*, GPT) and masked token prediction (*e.g.*, BERT) have led to highly successful unsupervised pretraining paradigms. The discrete nature of language makes these objectives well-defined and effective for large-scale text modeling. In contrast, designing effective pretext tasks in the CV domain has proven more challenging due to the continuous nature of visual



(a) SignDINO Pretraining **(b) SignDINO Finetuning**

Figure 2. The SignDINO Architecture with sign-aware DINO training strategy. We omit the global views of the student model for clarity.

features. Nevertheless, recent advances such as Masked Autoencoders (MAE) [20] and contrastive learning methods (*e.g.*, BYOL [14], MoCo [19], and SimSiam [3]) have demonstrated the feasibility of both generative and discriminative SSL approaches for visual data. Building upon these foundations, DINOs [39] adopt multiple global views of an image and introduce a self-distillation framework without labels, showing that strong and robust visual representations can be learned purely through self-supervision.

Building upon these advances, one may ask: **Can the rapidly evolving SSL techniques be directly applied to the GFSLT task, serving as a powerful SL tokenizer pretraining strategy without the need for any gloss or text supervision?**

Sign Language Tasks. Following the development of CSLR models [11, 30, 31, 43], SLT methods [33, 46] commonly follow the paradigm of first pretraining a CSLR backbone [6, 22, 35] and then using a translation model to generate textual sequences. Gloss-based SLT approaches [9, 24, 42] rely heavily on gloss annotations to achieve temporal segmentation and semantic alignment. Consequently, a CTC loss is typically used to pretrain the SL tokenizer before fine-tuning the translation module. These pretrained backbones effectively align sign video features with gloss sequences, serving as a form of "tokenizer" for sign videos.

Despite their superior performance, the strong dependence on gloss annotations severely limits the scalability and generalization of these models. Gloss-based SLT methods cannot be directly applied to datasets without gloss annotations such as OpenASL and How2Sign, motivating a growing interest in gloss-free SLT (GFSLT). In GFSLT, a key challenge lies in how to effectively tokenize sign videos and provide discriminative and semantically meaningful visual representations for translation models, which is crucial for finetuning. Expectedly, existing GFSLT models rely on text annotations during pretraining, typically adopting contrastive image-text learning [25, 45, 49], CTC-based pseudo gloss constraints [12], text-aligned vector quantization [13], or multi-stage pretraining schemes [17]. Recent works such as SHuBERT [16] and SignMusketeers [15] further finetune DINO encoders on local areas (*e.g.*, hands and face) as SL tokenizers, without the need for text. However, these methods are limited to local informative regions and fail to capture comprehensive sign-related features from global frames. In this paper, we aim to develop a simple yet effective training paradigm that tokenizes SL videos directly from global frames without relying on gloss or text.

3. Method

Overall Framework. The target of SLT is to learn the distribution mapping between SL video input $f = \{f_i\}_{i=1}^{\theta}$ with θ frames and text sequence $w = \{w_i\}_{i=1}^{\varsigma}$ with ς words: $p(w|f)$. For more detailed processing, an SL Tokenizer is adopted to extract SL features to get visual features $v = \mathcal{VE}(f)$ with m vectors, where $v = \{v_i \in \mathbb{R}^{d_T}\}_{i=1}^{m}$ ($m \leq \theta$) and d_T is the input dimension of the translation model. The text embedding module tokenizes the text sequence and gets n token ids $t = \{t_i\}_{i=1}^{n}$ ($n \geq \varsigma$). During training stage, the model predicts token $\hat{t}_i$ based on v and previous ground truth tokens: $p(\hat{t}_i|v, \{t_j\}_{j=0}^{i-1})$. While during testing stage, the model predicts token $\hat{t}_i$ based on v and previous predicted tokens: $p(\hat{t}_i|v, \{\hat{t}_j\}_{j=0}^{i-1})$.

3.1. The pretraining strategy for SLT

For gloss-based SLT models [4, 12], the pretext task used to train the visual backbone is typically the continuous sign language recognition (CSLR) task. Specifically, the SL tok-

enizer (*i.e.*, visual encoder) $\mathcal{VE}$ is trained under the Connectionist Temporal Classification (CTC) objective, where the gloss annotations g serve as supervision to optimize $\mathcal{VE}$ for learning temporal segmentation and semantic alignment:

$$\Theta^*_{\mathcal{VE},w} = \underset{\Theta_{\mathcal{VE},w}}{\arg\min}\ \mathbb{E}_{(f,g)\sim\mathcal{D}}\Big[\mathcal{L}_{\text{CTC}}(\mathcal{VE}(f)\cdot w, g)\Big], \quad (1)$$

where $\mathcal{L}_{CTC}$ denotes CSLR loss function, w denotes the classifier weights used in the CSLR task, $\mathcal{D}$ denotes the dataset. After training a well-performed CSLR model, A translation model $\mathcal{TR}$ (*e.g.*, mBART, T5, GPT-2), parameterized by $\Theta_{\mathcal{TR}}$, takes v (the output of $\mathcal{VE}$) as input and optimized under CE loss:

$$\Theta^*_{\mathcal{TR}} = \underset{\Theta_{\mathcal{TR}}}{\arg\min}\ \mathbb{E}_{(v,t)\sim\mathcal{D}}\Big[-\log p_{\Theta_{\mathcal{TR}}}(t \mid \mathcal{TR}(v))\Big] \quad (2)$$

For gloss-free SLT models, various pretext tasks are explored to pretrain the SL tokenizer, such as contrastive alignment with text and pseudo-gloss generation. In these methods, the text annotations t serve as supervision signals to optimize the SL Tokenizer $\mathcal{VE}$ for learning semantic alignment:

$$\Theta^*_{\mathcal{VE}}, \Theta^*_{\mathcal{TE}} = \underset{\Theta_{\mathcal{VE}},\Theta_{\mathcal{TE}}}{\arg\min}\ \mathbb{E}_{(f,t)\sim\mathcal{D}}\Big[\mathcal{L}_p(\mathcal{VE}(f), \mathcal{TE}(t))\Big], \quad (3)$$

where $\mathcal{L}_p$ denotes loss functions of the pretext tasks, $\mathcal{TE}$ denotes text processing functions (including: preprocessing, text embedding *etc*.)

For SLT, the pretrained backbone stage is critical to model performance, analogous to the alignment stage in vision-language models (VLMs). It ensures that the SL tokenizer provides effective and discriminative SL representations. Although effective, current pretraining paradigms are constrained by their reliance on gloss or text annotations.

3.2. Sign-aware DINO pretraining strategies

Following the DINO framework, we design a simple sign-aware DINO pretraining strategy. Specifically, we train two identical networks, a student encoder $\mathcal{VE}_s$ and a teacher encoder $\mathcal{VE}_t$, parameterized by Θ_s and Θ_t respectively, both sharing the same architecture. Each network is composed of an SL tokenizer followed by a DINO head, where the DINO head consists of several linear projection layers that map visual features into a common embedding space for self-distillation.

As shown in Figure 2, given an SL frame $x \in f$, we first construct two types of views: a global view set and a local view set. Instead of directly following the standard multi-crop augmentation strategy used in DINO, we introduce a sign-aware data augmentation scheme tailored for sign language videos. For the global view set, we generate two general global views, x^g_1 and x^g_2, using a series of standard data augmentations denoted as $\mathcal{DA}_g$. For the local view set, we construct three local views $\{x^l_i\}^3_{i=1}$, each capturing different discriminative regions: (1) the facial region, (2) the hand regions, and (3) the combination of face and hands. Importantly, we do not resize these local regions to the same scale as the global views. Instead, we retain their original spatial size and mask out all non-relevant areas. The student model takes both global views and local views, while only the global view is passed through the teacher model, and then we can get the probability distributions over K dimensions, denoted by P_s,

$$P_s(x)^j = \frac{exp(\mathcal{VE}_s(x)^j/\tau_s)}{\sum_{k=0}^{K} exp(\mathcal{VE}_s(x)^k/\tau_s)}. \quad (4)$$

The teacher distribution P_t is defined analogously. we optimize the student model $\mathcal{VE}_s$ as follows:

$$\Theta^*_{\mathcal{VE}_s} = \underset{\Theta_{\mathcal{VE}_s}}{\arg\min}\ \mathbb{E}_{x\sim\mathcal{D}}\left[\sum_{x\in x^g}\sum_{\substack{x'\in x^g, x^l \\ x'\neq x}} H(P_t(x), P_s(x'))\right], \quad (5)$$

where $H(x,y) = -x\log y$. The teacher parameters are updated as the exponential moving average (EMA) of the student's parameters.

$$\Theta_{\mathcal{VE}_t} \longleftarrow \lambda\Theta_{\mathcal{VE}_t} + (1-\lambda)\Theta_{\mathcal{VE}_s}, \quad (6)$$

where λ is a weight parameter following a cosine schedule from 0.996 to 1 during training.

3.3. Fine tuning for GFSLT

The final training stage begins by pretraining the SL tokenizer using the proposed sign-aware DINO strategy, as formulated in Equation 5. After pretraining, the teacher model weights are retained and integrated with the translation model to form the full GFSLT system. The combined model is then fine-tuned under the translation objective, using a standard cross-entropy loss, as shown in Equation 2.

During inference, the teacher model takes only global sign language frames as input without requiring any additional local regions to generate sign language features. The translation model then takes these features to produce the corresponding text sequence.

4. Experiments

Datasets. We evaluate our model on four widely used SLT datasets [40] commonly adopted in existing GFSLT studies: (1) Phoenix14T [1] is a German Sign Language (DGS) dataset containing 7,096 training, 519 validation, and 642 test samples from nine signers. It includes a vocabulary of 1,066 glosses and 2,877 German words. (2) CSL-Daily [50] is a Chinese Sign Language (CSL) dataset consisting of

Model	TEST				
	ROUGE	BLEU1	BLEU2	BLEU3	BLEU4
MAE ❄	17.81	17.38	7.23	5.17	4.19
DINO ❄	21.32	22.73	10.76	7.00	5.22
MAE♣	27.63	38.46	25.97	18.73	14.37
SimSiam♣	38.93	39.54	26.99	21.33	15.24
DINO♣	39.26	42.25	29.29	21.93	15.48
SignDINO	53.79	54.15	38.40	33.44	27.17

Table 1. Effect of current SSL pretraining strategies. ❄ indicates models initialized from HuggingFace weights without additional pretraining on PHOENIX14T. ♣ denotes backbones pretrained on PHOENIX14T using their respective SSL strategies.

18,401 training, 1,077 validation, and 1,176 test videos collected from ten signers. It provides 2,000 gloss tokens and 2,343 Chinese words for translation. (3) How2Sign [7] is a large-scale American Sign Language (ASL) dataset containing over 80 hours of multi-view recordings. We use only the RGB frontal-view videos, comprising 31,128 training, 1,741 validation, and 2,322 test samples. (4) OpenASL [38] is another large-scale ASL dataset with more than 280 hours of signing videos from over 200 signers. It includes 96,476 training, 997 validation, and 999 test samples.

Data Preprocessing. During pretraining, following the DINO series [2, 32, 39], we apply a data augmentation strategy $\mathcal{DA}_g$ that includes: resizing frames to 256×256 pixels, random cropping to 224×224 pixels, random horizontal flipping with a probability of 0.5, random grayscaling with a probability of 0.2, random Gaussian blurring with a probability of 0.1, and color jittering (probability 0.1) with parameters brightness=0.4, contrast=0.4, saturation=0.4, hue=0.1. During GFSLT fine-tuning, we adopt standard augmentations, including resizing to 256×256 pixels, random cropping to 224×224 pixels, random horizontal flipping (probability 0.5), and random temporal scaling within ±20%. During inference, frames are resized to 256×256 pixels and center-cropped to 224×224 pixels.

Implementation Details. Our architecture consists of three key components: (1) *Visual Encoder:* In the baseline setting, we adopt the DINOv3-Base model [39] as the visual backbone [11, 29]. Due to limited GPU resources, we employ LoRA fine-tuning instead of full model training to adapt the ViT-based backbone efficiently. (2) *DINO Head:* The DINO head consists of three MLP layers followed by an L2 normalization layer and another MLP projection layer. The output dim K of DINO head is set to 65536. (3) *Translation Model:* For a fair comparison with existing GFSLT studies, we adopt the widely used mbart architecture with a temporal convolution module [28] as our translation model in the baseline setting. Our model is implemented using PyTorch 2.6 and trained on four NVIDIA RTX 4090 GPUs (each with 24GB memory) in half-precision. It is important to note that our model is trained without using any additional SL data beyond the official training sets.

Data Augmentation			TEST				
Global	Face	Hands	ROUGE	BLEU1	BLEU2	BLEU3	BLEU4
—Baseline—			34.65	36.10	25.20	13.75	10.41
✓			39.26	42.25	29.29	21.93	15.48
	✓		28.06	28.64	16.24	10.81	7.74
		✓	40.15	38.64	26.96	20.11	15.79
✓	✓		39.63	41.78	28.91	21.50	16.92
✓		✓	49.63	51.78	36.91	29.50	25.92
	✓	✓	41.26	42.25	32.29	26.93	20.28
✓	✓	✓	53.79	54.15	38.40	33.44	27.17

Table 2. Effect of sign-aware pretraining.

Evaluation Metrics. Following prior GFSLT studies, we adopt ROUGE-L F1 [26], BLEU-1/2/3/4 [34] to evaluate GFSLT performance. These metrics are widely used in SLT research [4, 9].

5. Ablation Study

In this section, we aim to investigate several noteworthy issues through ablation studies on the PHOENIX14T dataset, as well as to verify the effectiveness of the proposed SignDINO model.

Q1: Can SOTA SSL methods be directly applied to the sign language domain directly? In this paper, we are interested in how existing self-supervised learning (SSL) methods perform when directly applied to the sign language domain. As shown in Table 1, we evaluate the effects of different SSL pretraining strategies. Specifically, we first directly apply SOTA SSL backbones pretrained on general-domain data and use them as frozen sign tokenizers during GFSLT training. We further pretrain the following backbones on the Phoenix14T dataset using their corresponding SSL paradigms, including generative masked image modeling (*i.e.*, MAE) and contrastive unsupervised learning (*i.e.*, SimSiam and DINO), and then finetune on GFSLT tasks.

As can be seen, directly adopting the DINO or MAE backbones as a feature extractor fails to yield satisfactory translation results. Moreover, simply applying existing SSL training strategies, such as MAE, SimSiam and DINO, to sign language videos also leads to suboptimal performance. The main reason is that these visual SSL algorithms are inclined to capture global semantic representations rather than focusing on fine-grained, discriminative local cues. In contrast, SignDINO, which explicitly emphasizes the discriminative local features of the face and hands, achieves the best performance, validating the effectiveness of our approach.

Q2: How do different sign-aware training strategies affect performance? As shown in Table 2, we evaluate different sign-aware data augmentation strategies for the student model, while keeping the teacher model fixed with a global view as input. Specifically, we vary the student's input views among global, face, and hand regions, as well as

Backbone	Parameters (M)		Test		
	Total	Trainable	ROUGE	BLEU1	BLEU4
PoolFormer-s12	12.0	12.0	53.01	53.52	26.97
SwinTransformer-s	48.8	48.8	52.75	53.03	25.60
DINO2-s	22.1	22.1	53.70	52.14	27.04
DINO3-s	21.6	21.6	52.76	53.01	26.87
DINO2-b (lora)	86.8	0.20	53.45	53.73	26.89
SwinTransformer-b (lora)	87.0	0.29	52.16	53.15	27.24
DINO3-b (lora)	85.8	0.20	53.79	54.15	27.17

Table 3. Effect of different backbones.

Init weights	TEST				
	ROUGE	BLEU1	BLEU2	BLEU3	BLEU4
DINO3-s (scratch)	51.46	51.16	37.41	32.02	25.01
DINO3-s	52.76	53.01	37.56	33.05	26.87

Table 4. Effect of weight initialization.

their different combinations. The results show that the baseline model (*i.e.*, trained in an end-to-end manner without SSL pretraining) performs the worst. Using only a single view (*i.e.*, global, face region, or hand regions) usually leads to inferior performance, indicating that each individual view alone fails to capture the complete signing semantics. Combining the global view with hand regions significantly improves translation quality, suggesting that hand motion provides crucial cues for sign understanding. Incorporating all three views achieves the best overall performance across all evaluation metrics.

Q3: Does the backbone architecture affect performance? As shown in Table 3, we evaluate several mainstream visual backbones, including PoolFormer, SwinTransformer, DINOv2, and DINOv3, with different model sizes. Due to GPU memory limitations, the base-sized models are trained under the LoRA setting; thus, we report both the total and trainable parameter counts for reference. The results indicate that different backbones, when trained with our sign-aware DINO strategy, achieve comparable performance. This suggests that the proposed training method is generally effective and does not heavily depend on a specific backbone architecture.

Q4: Does pretraining on other domains boost performance? As shown in Table 4, we investigate the effect of initialization by comparing models pretrained under different domains using the DINO-S backbone. (Note that DINO-B models can only be trained with LoRA due to GPU memory constraints and cannot be trained from scratch.) The original DINO models are pretrained on the LVD-1689M dataset (a collection of 1.689 billion images). We compare their performance with models trained entirely from scratch on SL data.

The results show that pretraining on other domains (*i.e.*, using the pretrained initialization weights) can improve performance. However, models trained from scratch also achieve satisfactory results, demonstrating that the performance gains primarily stem from our proposed sign-aware pretraining strategy rather than from large-scale pretraining on the LVD-1689M dataset.

Pretrain/Finetuning	CSL-Daily			Phoenix14T		
	ROUGE	BLEU1	BLEU4	ROUGE	BLEU1	BLEU4
CSL-daily	52.75	52.13	25.46	41.83	42.53	17.36
Phoenix14T	42.10	42.06	17.16	53.79	54.15	27.17

Table 5. Effect of cross dataset pretraining.

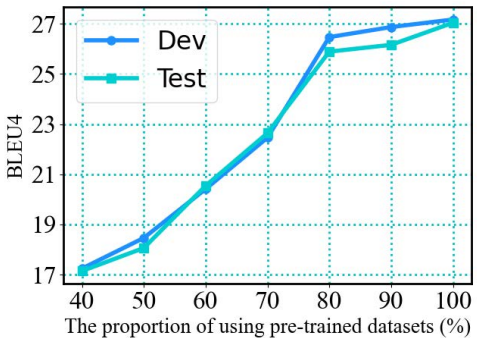


Figure 3. Scalability test: the effect of using different sizes of pretrained dataset.

Q5: How generalizable is the proposed sign-aware pretraining strategy? As shown in Table 5, we evaluate the generalization ability of the proposed sign-aware pretraining strategy by conducting cross-dataset experiments. Specifically, we pretrain the model on one dataset (*e.g.*, CSL-Daily) and then fine-tune it on another (*e.g.*, Phoenix14T), and vice versa. (Note that during fine-tuning, the backbone is frozen and only the translation module is trained.) The results demonstrate that models pretrained on one SL dataset can transfer to another. For instance, pretraining on CSL-Daily and fine-tuning on Phoenix14T still yields good results (ROUGE 41.83 vs. 53.79 under in-domain training), indicating that the learned representations are not overfitted to a specific dataset or signer distribution. This confirms that our sign-aware pretraining equips the model with transferable representational ability, demonstrating generalization across different SL datasets. As for the inferior performance compared to pretraining on the target dataset itself, the main reason may lie in the limited scale of the pretraining data. With larger and more diverse pretraining datasets, such overfitting can be alleviated.

Q6: How scalable is the proposed sign-aware pretraining strategy? As shown in Figure 3, we evaluate the scalability of SignDINO by pretraining the backbone on different proportions of the Phoenix14T and then fine-tuning it for the GFSLT task. The results show a consistent improvement in BLEU-4 performance as the proportion of pretraining data increases, demonstrating the good scalability of our

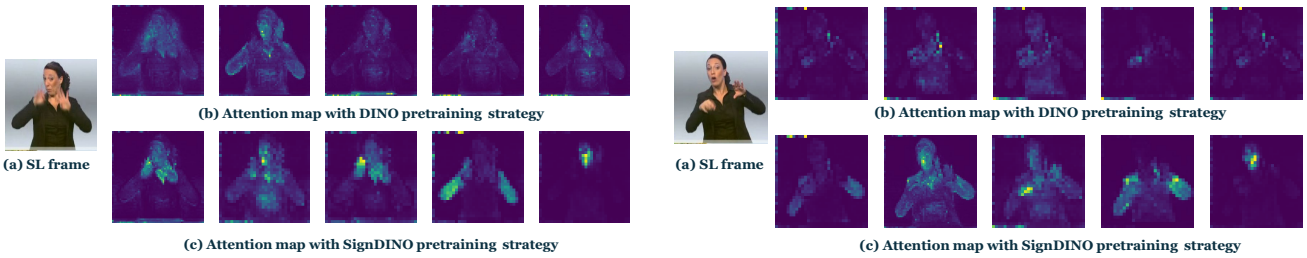


Figure 4. Visualization of attention map in SL tokenizer trained with original DINO/our SignDINO strategy.

Gloss free SLT	Extra Input			Phoenix14T									
				DEV					TEST				
	MT	Pose	Text	ROUGE	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE	BLEU1	BLEU2	BLEU3	BLEU4
PGG-SLT [17]	✓		✓	52.01	53.16	40.38	32.66	27.09	51.85	53.45	40.55	32.65	26.85
GFSLT-VLP [47]			✓	43.72	44.08	33.56	26.74	22.12	42.49	43.71	33.18	26.11	21.44
SignCL [45]			✓						49.04	49.76	36.85	29.97	22.74
Sign2GPT[44]			✓	-	-	-	-	-	48.90	49.54	35.96	28.83	22.52
GFSLT-VLP-SignCL [45]			✓	-	-	-	-	-	49.04	49.76	36.85	29.97	22.74
SignLLM [13]			✓	44.49	46.88	36.59	29.91	25.25	47.23	45.21	34.78	28.05	23.40
LLaVA-SLT [25]			✓	-	-	-	-	-	50.44	51.20	37.51	29.39	23.43
FLa-LLM [5]			✓	-	-	-	-	-	45.27	46.29	35.33	28.03	23.09
C2RL [6]			✓	-	-	-	-	-	50.96	52.81	40.20	32.20	26.75
MixSignGraph [12]			✓	51.71	51.07	37.97	29.98	24.87	51.14	50.01	38.04	29.95	24.02
SignDINO				53.61	53.49	38.40	33.01	27.17	53.79	54.15	38.40	33.44	27.17

Table 6. Comparison of SLT performance on Phoenix14T dataset. 'MT' denotes the multi-stage pretraining. 'Pose' denotes pose input. 'Text' denotes pretraining backbone with text annotation. Same applies to the tables below.

method. This suggests that further enlarging the pretraining dataset may lead to additional performance gains.

Training and Inference Speed. As shown in Table 7 and Table 8, we report the training and inference efficiency of our SignDINO model. The experiments are conducted on four NVIDIA RTX 4090 GPUs using the Phoenix14T dataset, where each video contains an average of 250 frames. The inference speed is measured on a single RTX 4090 GPU and averaged over 100 runs for estimation. For pretraining, the model takes approximately 30 minutes per epoch, while fine-tuning requires around 24 minutes per epoch. During inference, our model processes about 1.5 videos per second (each with 250 frames) on a single GPU. These results show that the proposed SignDINO achieves good efficiency in both training and inference, making it practical for large-scale SL learning.

6. Qualitative Results

As shown in Figure 4, we visualize the attention maps from multiple heads of the last-layer backbone trained with the original DINO and our proposed SignDINO pretraining strategy, respectively. Each visualization depicts the self-attention distribution of the [CLS] token using 224×224 input images. We observe that the original DINO pretraining strategy fails to effectively attend to the discriminative local regions (*e.g.*, face and hands), focusing mainly on the global human silhouette rather than fine-grained features. In contrast, our sign-aware DINO training strategy encourages the model to better capture the discriminative local cues solely based on the global frame input, demonstrating its effectiveness in the sign language domain. *More visualization can be found in the Appendix.*

Model	Training Time
Pretraining	30min/epoch
Finetuning	24min/epoch

Table 7. Model training speed.

Frame	Inference Time
250 frames	1.5 video/s

Table 8. Inference speed.

7. Comparisons

For fair comparison, we only include gloss-free SLT methods that do not leverage any external sign language datasets for pretraining. Consequently, gloss-based SLT models and those utilizing large-scale sign language datasets for pretraining, such as Scaling [48], Sign-Musketeers [15], and SHuBERT [16] *etc.*, are excluded from our comparison. *Please refer to the Appendix for more results on other SL-related tasks.*

Evaluation on Phoenix14T Dataset. As shown in Table 6, we compare our model with existing GFSLT methods and report the performance on both the validation and test sets. Current approaches typically rely on text-based pretraining of their backbones, including contrastive pretraining with text (*e.g.*, GFSLT-VLP-SignCL, C2RL, LLaVA-SLT) or CTC-constrained pretraining with pseudo glosses derived from text (*e.g.*, MixSignGraph and PGG-SLT). Our pro-

Gloss free SLT	Extra Input			CSL-Daily									
				DEV					TEST				
	MT	Pose	Text	ROUGE	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE	BLEU1	BLEU2	BLEU3	BLEU4
PGG-SLT(mBART) [17]	✓		✓	53.58	52.34	-	-	24.05	53.29	52.88	-	-	23.70
GFSLT-VLP [49]			✓	36.44	39.20	25.02	16.35	11.07	36.70	39.37	24.93	16.26	11.00
SignCL [45]			✓						48.92	47.47	32.53	22.62	16.16
Sign2GPT [44]			✓	-	-	-	-	-	42.36	41.75	28.73	20.60	15.40
SignLLM [13]			✓	39.18	42.45	26.88	17.90	12.23	39.91	39.55	28.13	20.07	15.75
GFSLT-VLP-SignCL [45]			✓	-	-	-	-	-	48.92	47.47	32.53	22.62	16.16
FLa-LLM [5]			✓	-	-	-	-	-	37.25	37.13	25.12	18.38	14.20
C^2RL [6]			✓	-	-	-	-	-	48.21	49.32	36.28	27.54	21.61
LLaVA-SLT [25]			✓	-	-	-	-	-	51.26	52.15	36.24	26.47	20.42
MixSignGraph [12]			✓	49.16	49.98	36.42	26.89	20.43	49.93	50.24	36.91	27.54	20.78
SignDINO				52.36	53.64	38.65	30.49	25.62	52.75	52.13	39.64	33.73	25.46

Table 9. Comparison of GFSLT performance on CSL-Daily.

Gloss free SLT	Extra Input			OpenASL									
				DEV					TEST				
	MT	Pose	Text	ROUGE	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE	BLEU1	BLEU2	BLEU3	BLEU4
GloFE-VN [27]			✓	21.37	21.06	12.34	8.68	6.68	21.75	21.56	12.74	9.05	7.06
C^2RL [6]			✓	-	-	-	-	-	31.36	31.46	21.85	16.58	13.21
MixSignGraph [12]			✓	25.41	26.82	16.70	11.48	8.36	25.71	26.65	16.55	11.68	8.69
Conv-GRU [38]				16.25	16.72	8.95	6.31	4.82	16.10	16.11	8.85	6.18	4.58
I3D-transformer [38]				18.88	18.26	10.26	7.17	5.60	18.64	18.31	10.15	7.19	5.66
OpenASL [38]				25.31	24.35	14.94	10.72	8.39	24.83	23.87	14.08	9.90	7.54
SignDINO				37.66	36.14	27.46	21.58	16.89	38.65	38.64	26.79	18.62	17.46

Table 10. Comparison of GFSLT performance on OpenASL.

Gloss free SLT	Extra Input			TEST				
	MT	Pose	Text	ROUGE	BLEU1	BLEU2	BLEU3	BLEU4
PGG-SLT(mBART) [17]	✓		✓	31.50	38.90	25.40	18.10	13.10
YouTube-SLT* [42]			✓	-	14.96	5.11	2.26	1.22
C^2RL [6]			✓	27.02	29.07	18.56	12.92	9.37
FLa-LLM [5]			✓	27.81	29.81	18.99	13.27	9.66
GloFE-VN [27]			✓	12.61	14.94	7.27	3.93	2.24
SSVP-SLT [37]			✓	25.70	30.20	16.70	10.50	7.00
MixSignGraph [12]			✓	28.01	34.74	20.83	14.41	10.41
SLT-IV [41]				-	34.01	19.30	12.18	8.03
SignDINO				36.14	40.51	26.95	16.40	15.47

Table 11. Comparison of GFSLT performance on How2Sign.

posed sign-aware pretraining strategy removes the dependency on text-based supervision and directly models visual structures relevant to sign language, such as hand and facial movements. As shown, our model even outperforms the most text-based pretraining strategies across multiple evaluation metrics, demonstrating the effectiveness of our visual-centric learning paradigm for the GFSLT task.

Evaluation on CSL-daily Dataset. We also show the GFSLT performance of our model on CSL-daily dataset. Table 9 shows that our model achieves 52.36, 52.75 ROUGE on the dev, test sets, respectively. It can be found that the proposed SignDINO can outperform most current text-based GFSLT models across multiple metrics.

Evaluation on OpenASL Dataset. OpenASL is a large-scale American Sign Language dataset that provides only text annotations. As shown in Table 10, SignDINO performs competitively across all evaluation metrics, even without relying on text-based pretraining. While methods such as MixSignGraph depend on pretraining with text, SignDINO attains comparable or better results by leveraging discriminative visual cues learned from global frames through the proposed sign-aware pretraining strategy.

Evaluation on How2Sign Dataset. Similar to OpenASL, How2Sign is a large-scale dataset with a diverse vocabulary and text-only annotations. Our model achieves strong performance on the test sets. As shown in Table 11, compared with the previous SOTA model: PGG-SLT, SignDINO improves ROUGE from 31.50 to 36.14 and BLEU-4 from 13.10 to 15.47 on the test set, indicating the effectiveness of the proposed sign-aware DINO pretraining strategy in large-scale SLT scenarios.

8. Conclusion

In this paper, we focus on designing a self-supervised pretrained strategy for sign language tasks, aiming to eliminate the dependency on gloss or text supervision during pretraining. We reveal that directly applying existing self-supervised learning (SSL) strategies to sign language datasets often fails to achieve satisfactory results. To address this issue, we propose a sign-aware DINO pretraining strategy, which feeds discriminative key local regions into the student model, enabling it to focus on sign-related features. Extensive experiments on multiple benchmark datasets demonstrate the effectiveness of our approach, with our model achieving very competitive performance on the gloss-free SLT task on multiple datasets.

9. Acknowledgments

This work is supported in part by National Natural Science Foundation of China under Grant Nos. 62172208, 92467202, 62272216; Key Projects of Jiangsu Provincial Basic Research Program under Grant No. BK20243040; JiangSu Natural Science Foundation under Grant No. BK20251989. This work is partially supported by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM118); Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

[1] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *CVPR*, pages 7784–7793, 2018. 4

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 5

[3] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021. 3

[4] Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems*, 35:17043–17056, 2022. 1, 3, 5, 13

[5] Zhigang Chen, Benjia Zhou, Jun Li, Jun Wan, Zhen Lei, Ning Jiang, Quan Lu, and Guoqing Zhao. Factorized learning assisted with large language model for gloss-free sign language translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7071–7081, 2024. 7, 8

[6] Zhigang Chen, Benjia Zhou, Yiqing Huang, Jun Wan, Yibo Hu, Hailin Shi, Yanyan Liang, Zhen Lei, and Du Zhang. C 2 rl: Content and context representation learning for gloss-free sign language translation and retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 3, 7, 8

[7] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2sign: a large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2735–2744, 2021. 5

[8] Edward Fish and Richard Bowden. Geo-sign: Hyperbolic contrastive regularisation for geometrically aware sign language translation. *arXiv preprint arXiv:2506.00129*, 2025. 1

[9] Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Lei Xie, and Sanglu Lu. Skeleton-aware neural sign language translation. In *MM*, pages 4353–4361, 2021. 3, 5

[10] Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Kang Xia, Lei Xie, and Sanglu Lu. 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, 2023. 1, 13

[11] Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Hongkai Wen, Lei Xie, and Sanglu Lu. Signgraph: A sign sequence is worth graphs of nodes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13470–13479, 2024. 3, 5, 13

[12] Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Lei Xie, Sanglu Lu, and Hongkai Wen. Mixsigngraph: A sign sequence is worth mixed graphs of nodes, 2025. 1, 3, 7, 8, 13

[13] Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. Llms are good sign language translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18362–18372, 2024. 1, 3, 7, 8

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NIPS*, 33:21271–21284, 2020. 3

[15] Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, and Karen Livescu. Signmusketeers: An efficient multi-stream approach for sign language translation at scale. *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. 2, 3, 7

[16] Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, Karen Livescu, and Alexander H. Liu. SHuBERT: Self-supervised sign language representation learning via multi-stream cluster prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28792–28810, Vienna, Austria, 2025. Association for Computational Linguistics. 1, 2, 3, 7

[17] Jianyuan Guo, Peike Li, and Trevor Cohn. Bridging sign and spoken languages: Pseudo gloss generation for sign language translation. *arXiv preprint arXiv:2505.15438*, 2025. 3, 7, 8

[18] Aiming Hao, Yuecong Min, and Xilin Chen. Self-mutual distillation learning for continuous sign language recognition. In *ICCV*, pages 11303–11312, 2021. 13

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 3

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 3

[21] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Continuous sign language recognition with correlation network. In *CVPR*, pages 2529–2539, 2023. 13

[22] Peiqi Jiao, Yuecong Min, Yanan Li, Xiaotao Wang, Lei Lei, and Xilin Chen. Cosign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition. In *ICCV*, pages 20676–20686, 2023. 3, 13

[23] Jichao Kan, Kun Hu, Markus Hagenbuchner, Ah Chung Tsoi, Mohammed Bennamoun, and Zhiyong Wang. 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, 2022. 13

[24] Dongxu Li, Chenchen Xu, Xin Yu, Kaihao Zhang, Benjamin Swift, Hanna Suominen, and Hongdong Li. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. In *NIPS*, 2020. 3

[25] Han Liang, Chengyu Huang, Yuecheng Xu, Cheng Tang, Weicai Ye, Juze Zhang, Xin Chen, Jingyi Yu, and Lan Xu. Llava-slt: Visual language tuning for sign language translation. *arXiv preprint arXiv:2412.16524*, 2024. 1, 3, 7, 8

[26] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5

[27] Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Yi Yang, et al. Gloss-free end-to-end sign language translation. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023. 8

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 5

[29] Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen. Visual alignment constraint for continuous sign language recognition. In *ICCV*, pages 11542–11551, 2021. 5, 13

[30] Yuecong Min, Yifan Yang, Peiqi Jiao, Zixi Nan, and Xilin Chen. A closer look at skeleton-based continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4909–4915, 2025. 3

[31] Zhe Niu and Brian Mak. Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. In *ECCV*, pages 172–186. Springer, 2020. 3

[32] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 2, 5

[33] Alptekin Orbay and Lale Akarun. Neural sign language translation by learning tokenization. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 222–228. IEEE, 2020. 3

[34] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002. 5

[35] Maria Parelli, Katerina Papadimitriou, Gerasimos Potamianos, Georgios Pavlakos, and Petros Maragos. Spatio-temporal graph convolutional networks for continuous sign language recognition. In *ICASSP*. IEEE, 2022. 3

[36] Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgöz, and Jean Maillard. Towards privacy-aware sign language translation at scale. *arXiv preprint arXiv:2402.09611*, 2024. 1

[37] Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgoz, and Jean Maillard. Towards privacy-aware sign language translation at scale. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8624–8641, Bangkok, Thailand, 2024. Association for Computational Linguistics. 8

[38] Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. Open-domain sign language translation learned from online video. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6365–6379, 2022. 5, 8

[39] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. 2, 3, 5

[40] Garrett Tanzer and Biao Zhang. Youtube-sl-25: A large-scale, open-domain multilingual sign language parallel corpus. *arXiv preprint arXiv:2407.11144*, 2024. 4

[41] Laia Tarrés, Gerard I Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró-i Nieto. Sign language translation from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5624–5634, 2023. 8

[42] Dave Uthus, Garrett Tanzer, and Manfred Georg. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. *Advances in Neural Information Processing Systems*, 36:29029–29047, 2023. 3, 8

[43] Fangyun Wei and Yutong Chen. Improving continuous sign language recognition with cross-lingual signs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23612–23621, 2023. 3

[44] Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. Sign2gpt: leveraging large language models for gloss-free sign language translation. In *ICLR 2024: The Twelfth International Conference on Learning Representations*, 2024. 1, 7, 8

[45] Jinhui Ye, Xing Wang, Wenxiang Jiao, Junwei Liang, and Hui Xiong. Improving gloss-free sign language translation by reducing representation density. In *NeurIPS*, 2024. 1, 3, 7, 8

[46] Aoxiong Yin, Zhou Zhao, Jinglin Liu, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. Simulslt: End-to-end simultaneous sign language translation. In *MM*, pages 4118–4127, 2021. 3

[47] Aoxiong Yin, Tianyun Zhong, Li Tang, Weike Jin, Tao Jin, and Zhou Zhao. Gloss attention for gloss-free sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2551–2562, 2023. 7

[48] Biao Zhang, Garrett Tanzer, and Orhan Firat. Scaling sign language translation. *Advances in neural information processing systems*, 37:114018–114047, 2024. 1, 7

[49] Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20871–20881, 2023. 3, 8, 13