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

10. Limitations and Discussion

Limitations. We outline several potential directions for future improvement. Although our SignDINO model takes only global frames as input during inference, the proposed sign-aware DINO training strategy requires local regions (e.g., face and hand areas) during training, which introduces additional preprocessing steps to obtain these region locations. Developing a simpler and more efficient self-supervised learning (SSL) strategy that avoids such extra processing during training is a promising direction for future work. Moreover, we acknowledge potential negative social impacts. As our method is data-driven, it may inherit biases present in the training data. Ensuring careful dataset selection and maintaining balanced data distribution are crucial to mitigate such biases.

Scaling to Larger Datasets. SignDINO removes the dependency on text-supervised pretraining of the SL tokenizer (*i.e.*, visual backbone), enabling the model to scale more effectively across diverse SL datasets, **including unlabeled raw SL videos**. Expanding the pretraining on larger and more diverse SL videos is expected to further enhance performance. However, due to current computational limitations, this work is restricted to moderate-scale datasets and backbones. Future research will explore scaling SignDINO to larger datasets and larger backbone sizes.

11. Training Details

Here, we describe the pretraining and fine-tuning details of SignDINO.

Pretraining. We adopt the AdamW optimizer with a batch size of 300 distributed over 4 GPUs when using ViT-S/16 (each GPU processes 100 images). The learning rate is set to 0.0005 and follows a cosine decay schedule. The weight decay also follows a cosine schedule from 0.04 to 0.4. The temperature τ_s is fixed at 0.1, while τ_t is linearly warmed up from 0.04 to 0.07 during the first 30 epochs. The total number of pretraining epochs is 100. The face and hand regions are cropped based on the keypoint coordinates extracted by HRNet, with a fixed crop size of 45×45 applied to each region. The outdim K of DINO head is set to 65536.

Fine-tuning. During fine-tuning, we follow previous works and adopt the mbart architecture. The model is fine-tuned for 40 epochs with an initial learning rate of 1×10^{-5}, which decays by a factor of 0.5 every 10 epochs.

Evaluation. For fair comparison, we report the model performance based on the best results on the development set, following prior work. The random seed is fixed to 0 for all experiments.

12. Visualization

We further visualize the attention maps of the sign language tokenizer trained with the original DINO and our proposed SignDINO strategies. We consider four representative cases: (1) frames containing only the upper body without hands, (2) frames with both hands visible, (3) frames with only one hand visible, and (4) frames with blurred hands. For each case, we present the ground-truth frame, the attention map generated by the original DINO, and the attention map generated by SignDINO. As shown in Figure 5, our SignDINO training strategy effectively focuses on discriminative local regions such as hands and face, while the original DINO tends to attend to the global human silhouette and fails to capture fine-grained hand-related cues.

13. Effect on Other SL Tasks

We further evaluate the effectiveness of our SignDINO model on the Continuous Sign Language Recognition (CSLR) task. After pretraining with our sign-aware DINO strategy, we integrate the pretrained backbone into a widely used CSLR framework, namely the VAC model, by replacing its original ResNet encoder with our SignDINO backbone. The entire model is then fine-tuned on the CSLR datasets.

As shown in Table 12, we evaluate the CSLR performance under two settings: (1) the backbone is frozen (*i.e.*, SignDINO❄), and (2) the backbone is trainable (*i.e.*, SignDINO). It can be observed that even when the backbone is frozen, our model still achieves competitive CSLR performance, indicating that the proposed pretraining strategy effectively learns discriminative sign representations. When the backbone is jointly fine-tuned, our model achieves superior performance compared to existing CSLR approaches on multiple datasets, even without relying on any additional cues. These results demonstrate that the proposed sign-aware pretraining strategy generalizes well beyond translation tasks.

For gloss-based SLT, we do not include evaluation results because these models require retraining the backbone with gloss annotations after our sign-aware pretraining. This process would alter the pretrained parameters and

Model	CSL-Daily				Phoenix				Phoenix14T	
	DEV		TEST		DEV		TEST		DEV	TEST
	WER	del/ins	WER	del/ins	WER	del/ins	WER	del/ins	WER	WER
HST-GNN [23]*	-	-	-	-	19.5	-/-	19.8	**-/-**	19.5	19.8
CoSign [22]*	28.1	-/-	27.2	-/-	19.7	-/-	20.1	**-/-**	19.5	20.1
TwoStream [4]*	25.4	-/-	25.3	-/-	18.4	-/-	18.8	-/-	17.7	19.3
C2SLR [51]	30.6	-/-	30.1	-/-	20.5	-/-	20.4	-/-	20.2	20.4
TwoStream [4]	28.9	-/-	28.5	-/-	22.4	-/-	23.3	-/-	21.1	22.4
SMKD [18]	-	-	-	-	20.8	6.8/2.5	21.0	6.3/2.3	20.8	22.4
CorrNet [21]	-	-	-	-	18.8	5.6/2.8	19.4	5.7/2.3	18.9	20.5
SignGraph [11]	27.3	7.9/2.3	26.4	7.8/2.1	18.2	4.9/2.0	19.1	5.3/1.9	17.8	19.1
Contrastive [10]	-	-	-	-	19.6	5.1/2.7	19.8	5.8/3.0	20.0	20.1
VAC [29] (baseline)	-	-	-	-	21.2	7.9/2.5	22.3	8.4/2.6	-	-
SignDINO❄	33.3	7.5/6.2	35.1	15.1/5.1	25.0	14.0/2.9	26.1	10.3/3.0	25.8	26.2
SignDINO	25.3	6.3/1.7	25.1	5.1/1.9	18.0	4.9/2.0	18.4	5.3/1.9	17.9	18.9

Table 12. Comparison of CSLR performance on CSL-Daily, Phoenix14 and Phoenix14T datasets. ($*$: using extra cues. ❄ means that the backbone is frozen.)

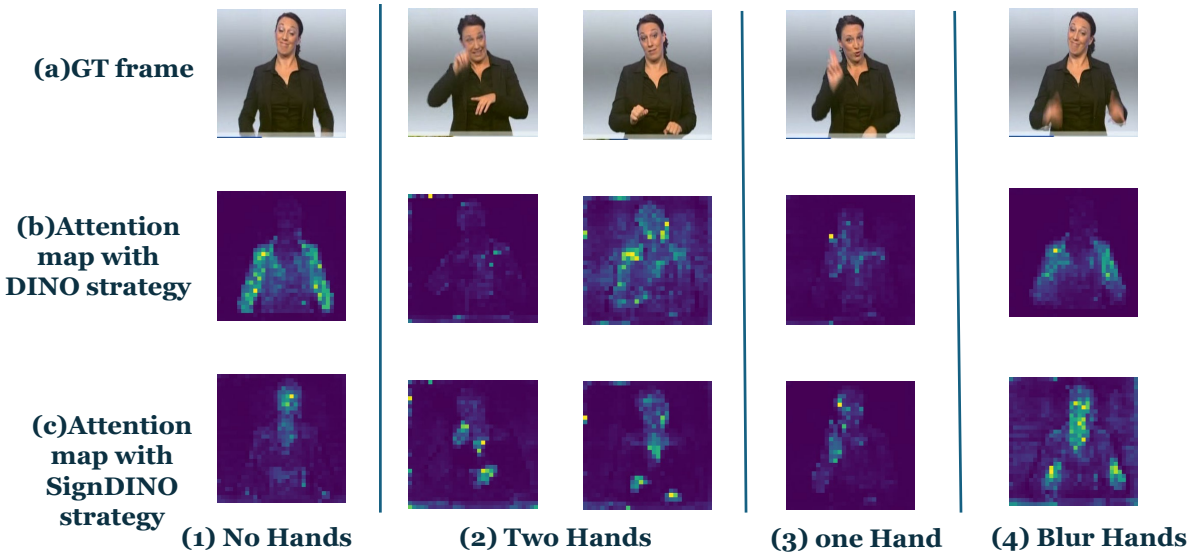

Figure 5. More visualization of attention map in SL tokenizer trained with original DINO/our signDINO strategy.

Pretrain/Finetuning	CSL-daily /Phoenix14t			Phoenix14t / CSL-daily		
	ROUGE	BLEU1	BLEU4	ROUGE	BLEU1	BLEU4
GFSLT-VLP [49]	30.14	27.02	10.47	30.56	31.78	9.81
MixSignGraph [12]	36.23	36.34	14.36	36.75	37.14	14.04
SignDINO	41.83	42.53	17.36	42.10	42.06	17.16

Table 13. Comparison of cross dataset setting.

thus make it difficult to fairly assess the effectiveness of our proposed training strategy.

14. GFSLT Qualitative Results

Table 14 presents several GFSLT translation examples on Phoenix14T, How2Sign, and OpenASL. Results on CSL-Daily are omitted because Chinese characters are not supported by the CVPR format. As shown in the examples, our SignDINO with the sign-aware training strategy produces the most accurate translations. In contrast, the original DINO pretraining strategy generates less satisfactory sentences, highlighting the effectiveness of the proposed sign-aware DINO training strategy for GFSLT.

15. Comparison under Cross-dataset Setting

To evaluate the effectiveness of our method under cross-dataset settings, as shown in Table 13, we compare SignDINO with two open-source baselines, MixSignGraph (pseudo-gloss-based) and GFSLT-VLP (contrastive-based), under cross-dataset pretraining. All baseline results are reproduced by ourselves for a fair comparison. SignDINO consistently outperforms both baselines, demonstrating its superior cross-domain generalization ability.

example(a)	Phoenix14T dataset
Groundtruth	im westen und nordwesten beruhigt sich das wetter später wieder . (In the west and northwest, the weather will calm down again later.)
DINO pretraining	im westen beruhigt sich das wetter später wieder . (The weather will calm down again later in the west.)
SignDINO	im westen und nordwesten beruhigt sich das wetter später wieder . (In the west and northwest, the weather will calm down again later.)
Groundtruth	am sonntag im norden und in der mitte schauer dabei ist es im norden stürmisch . (Showers are expected in the north and central regions on Sunday, with stormy conditions in the north.)
DINO pretraining	am Sonntag werden im norden schauer erwartet, besonders in küstennähe. (Showers are expected in the north on Sunday, especially near the coast.)
SignDINO	im norden und in der mitte schauer dabei ist es im norden stürmisch . (Showers in the north and center, with stormy conditions in the north.)
example(b)	How2Sign dataset
Groundtruth	We press this button, which opens the door.
DINO pretraining	We open the door.
SignDINO	We press this button, which will open the door.
Groundtruth	I'm sharing with you some secrets and tips on how to get your best performance from ironing.
DINO pretraining	I'm telling you tips about ironing.
SignDINO	I'm sharing with you some secrets and tips on how to get your best performance from ironing.
example(c)	OpenASL dataset
Groundtruth	Not to worry, we'll take good care of them and have fun!
DINO pretraining	You can leave, we can take care of ourselves.
SignDINO	Don't worry, we'll take good care of them and have fun!
Groundtruth	The virus attack rate has lowered as has the demand for supplies, such as hospital beds.
DINO pretraining	The virus rate is lower, as the bed demand decreases.
SignDINO	The virus attack rate has lowered as has the demand for supplies, such as beds.

Table 14. Translation Examples of GFSLT on Phoenix14T, How2Sign, and OpenASL.