– Supplementary Material – EvSign: Sign Language Recognition and Translation with Streaming Events

1 Translation Head

Following [1], we introduce an auto-regressive transformer decoder as our translation head, which contains four decoder blocks. Each decoder block consists of a masked self-attention layer, an encoder-decoder attention layer and a feed-forward layer. The spoken language sentence is first prefixed with a special beginning-of-sentence token (<bos>). Then, the generated word tokens are sent to the masked self-attention layer, where each token can only use its predecessors to extract contextual representation. The encoder-decoder attention layer is used to learn the mapping between gloss-aware and word tokens. Finally, a feed-forward layer is appended to predict the probability of words in spoken language. The translation head learns to generate target sentence in an auto-regressive manner until it produces a special end-of-sentences (<eos>).

2 Comparison between Synthetic and Real Data

As shown in Fig. 1, we claim that there is a gap between synthetic and real events. We conclude the advantages of the collected real event dataset EvSign from three aspects.

First, the synthetic events from RGB frames suffer from poor continuity, which relates to the framerate of RGB videos. In contrast, our dataset, captured by high-quality event cameras, can generate rich events within mircosecond response, thus providing smoother trajectories. Second, as illustrated in the second row, the RGB frames are blurry in fast motion scenarios. The blurriness leads to the loss of informative boundaries in event data, which are crucial cues for sign language tasks. However, the event frames sampled in real events contain sharper boundaries, which can provide more discriminative details, achieving better recognition results. As shown in Sec. 5 of the main submission, methods utilizing real events yield superior performance compared to those relying on RGB inputs. This shows the effectiveness of real events in sign language recognition and translation tasks. Third, we utilize an event-based image reconstruction method (E2VID [6]) to reconstruct intensity images from events, shown in the third row of Fig. 1. Given little movement with signer's head, the reconstruction method cannot recover signer's many facial details, thereby protecting the user's privacy.

3 Comparison on SL-Animals-DVS

Since EvASL is not publicly available, we conduct comparison on SL-Animals-DVS for ISLR. Please note that we focus on CSLR, which is different from ISLR.

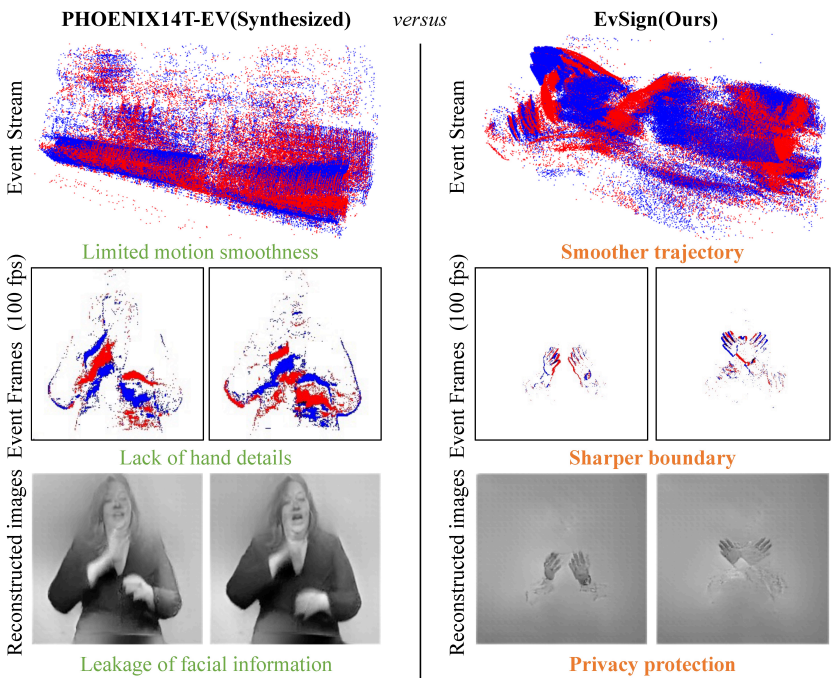

Fig. 1: Comparison between synthetic and real Data.

Thus, we modify those methods by replacing CTC loss with cross-entropy loss and aggregating features along temporal dimension to perform sequence-level prediction rather than frame-level prediction. Following previous methods, we adopt 4-fold cross-validation and report the mean and standard deviation of accuracy (Acc.) in Tab. 1. Our method outperforms other methods with 3.45% improvement.

Table 1: Results on SL-Animals-DVS dataset.

	VAC	TLP	SEN	CorrNet	Ours
Acc.(%)	95.48±1.15	96.23±0.46	96.05±0.82	95.38±1.10	**99.68±0.09**

4 More Results on Sign Language Recognition

We also conduct a comparison of sign language recognition (SLR) task on the synthetic CSL-Daily dataset, shown in Table 2. We note that the synthetic events

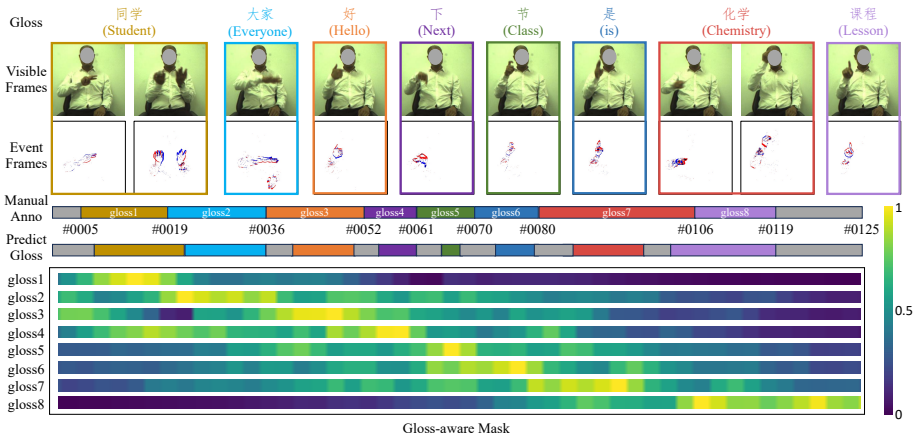

Fig. 2: Comparison between synthetic and real Data.

do not bring gains to SLR methods due to the limited smoothness and blurry content with the original frame sequences. This also reflects the necessity for proposing a high-quality sign language dataset with real events. Compared to the state-of-the-art method (CorrNet [3]), our method achieves 0.81%/0.88% WER reduction in both Dev and Test subsets, which demonstrates its effectiveness in handling sign language recognition.

Table 2: Comparison results for SLR on CSL-Daily dataset.

Modal	RGB				EV				
Method	VAC [5]	TLP [2]	SEN [4]	CorrNet [3]	VAC [5]	TLP [2]	SEN [4]	CorrNet [3]	Ours_{S2G}
Dev(%)	31.24	32.30	31.10	30.60	35.16	36.10	35.42	<u>34.81</u>	**34.00**
Test(%)	30.68	32.35	30.70	30.10	34.75	36.18	35.27	<u>34.70</u>	**33.82**

5 More Visualization of Gloss-Aware Mask

In this section, we aim to conduct qualitative analysis on the proposed Gloss-Aware Mask Attention (GAMA). Fig. 2 illustrates the learned mask in GAMA. We claim that it is capable of adaptively aggregating tokens that belong to the same gloss, thus modeling the complete sign language gestures and eliminating the interference from different glosses.

6 Analysis on weight of loss components

Following VAC and CorrNet, we set all the weights to 1 without modification. We conduct analysis on EvSign to evaluate the influence of loss weights. As shown in Tab. 3, our method is not sensitive to variations in loss weights.

Table 3: Analysis of loss weights on EvSign dataset.

λ_{inter}	1	1	1	5	5	10	10
λ_{final}	1	5	10	1	10	1	5
Dev (%)	**29.19**	29.43	29.27	29.33	29.24	29.59	30.17
Test (%)	28.69	29.37	29.01	29.07	**28.58**	28.96	30.22

References

1. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: Joint end-to-end sign language recognition and translation. In: CVPR (2020)
2. Hu, L., Gao, L., Liu, Z., Feng, W.: Temporal lift pooling for continuous sign language recognition. In: ECCV (2022)
3. Hu, L., Gao, L., Liu, Z., Feng, W.: Continuous sign language recognition with correlation network. In: CVPR (2023)
4. Hu, L., Gao, L., Liu, Z., Feng, W.: Self-emphasizing network for continuous sign language recognition. In: AAAI (2023)
5. Min, Y., Hao, A., Chai, X., Chen, X.: Visual alignment constraint for continuous sign language recognition. In: ICCV (2021)
6. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: High speed and high dynamic range video with an event camera. IEEE Trans. Pattern Anal. Mach. Intell. **43**(6), 1964–1980 (2021)