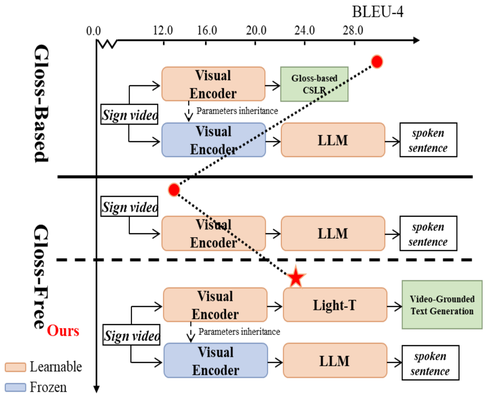
Factorized Learning Assisted with Large Language Model for Gloss-free Sign Language Translation

Zhigang Chen, Benjia Zhou, Jun Li, Jun Wan*, Zhen Lei, Ning Jiang, Quan Lu, Guoqing Zhao
* Corresponding Author

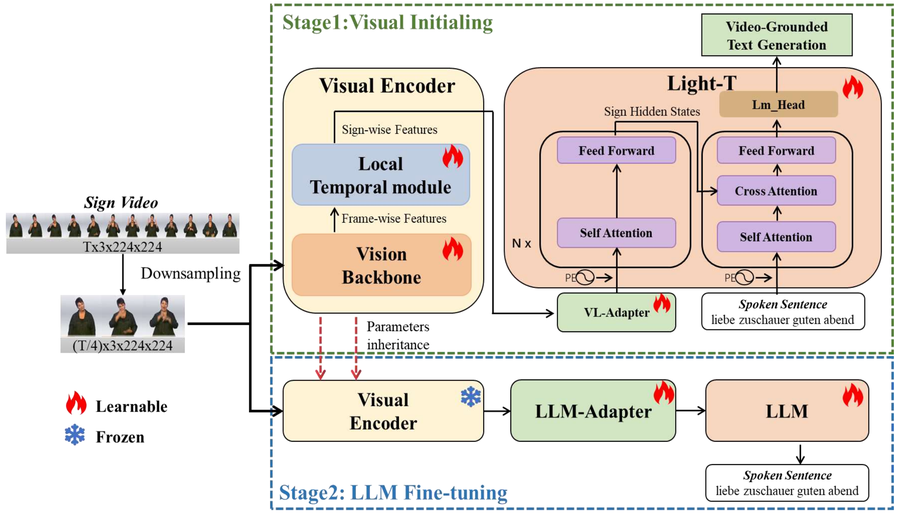
LREC-COLING 2024

Motivation



Observation:
1. LLM can improve gloss-based SLT.
2. Directly training the visual encoder and LLM failed in gloss-free SLT.

Method



Visual Initialing:
A visual encoder followed by a lightweight translation model (Light-T) to perform visual initialing by a video-grounded text generation task.

LLM Fine-tuning:
1. Freezing the Visual Encoder.
2. Fine-tuning with Mbart using sequence-to-sequence cross-entropy loss.

$$p(o_i|o_{1:i-1},V)=(\text{softmax}(\text{Lm\_Head}(y_i)))_{o_i}$$

$$\mathcal{L}_{CE}=-\sum_{i=1}^{L}\log p(o_i|o_{1:i-1},V).$$

Experiments

New **SOTA** has been achieved on three SLTdatasets

Method	Gloss-Free	Rouge-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
SLRT (Camgoz et al., 2020)	×	-	46.61	33.73	26.19	21.32
STMC-T (Zhou et al., 2021b)	×	46.65	46.98	36.09	28.70	23.65
SignBT (Zhou et al., 2021a)	×	49.54	50.80	37.75	29.72	24.32
MMTLB (Chen et al., 2022a)	×	52.65	53.97	41.75	33.84	28.39
TS-SLT (Chen et al., 2022b)	×	**53.48**	**54.90**	**42.43**	**34.46**	**28.95**
NSLT (Camgoz et al., 2018)	✓	31.80	32.24	19.03	12.83	9.58
SLRT-GF* (Camgoz et al., 2020)	✓	31.10	30.88	18.57	13.12	10.19
TK-SLT (Orbay and Akarun, 2020)	✓	36.28	37.22	23.88	17.08	13.25
TSPNet (Li et al., 2020)	✓	34.96	36.10	23.12	16.88	13.41
CSGCR (Zhao et al., 2021)	✓	38.85	36.71	25.40	18.86	15.18
GASLT (Yin et al., 2023)	✓	39.86	39.07	26.74	21.86	15.74
GFSLT-VLP (Zhou et al., 2023)	✓	42.49	43.71	33.18	26.11	21.44
FLa-LLM(ours)	✓	**45.27**	**46.29**	**35.33**	**28.03**	**23.09**
Improvement		+2.78	+2.58	+2.15	+1.92	+1.65

Table 1: Experimental results on PHOENIX14T dataset. * denotes methods reproduced by (Yin et al., 2023). We bold the best results in the gloss-based setting and gloss-free setting. Improvement represents comparisons with the previous best gloss-free result.

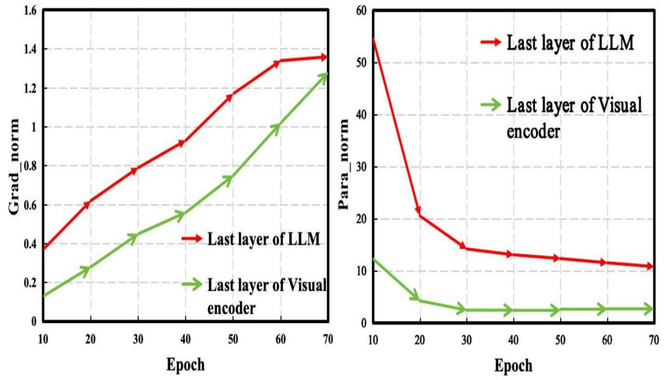
Method	Gloss-Free	Rouge-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
SLRT† (Camgoz et al., 2020)	×	36.74	37.38	24.36	16.55	11.79
SignBT (Zhou et al., 2021a)	×	49.31	51.42	37.26	27.76	21.34
MMTLB (Chen et al., 2022a)	×	53.25	53.31	40.41	30.87	23.92
TS-SLT (Chen et al., 2022b)	×	**55.72**	**55.44**	**42.59**	**32.87**	**25.79**
NSLT† (Camgoz et al., 2018)	✓	34.54	34.16	19.57	11.84	7.56
TSPNet* (Li et al., 2020)	✓	18.38	17.09	8.98	5.07	2.97
GASLT (Yin et al., 2023)	✓	20.35	19.90	9.94	5.98	4.07
GFSLT-VLP (Zhou et al., 2023)	✓	36.44	**39.37**	24.93	16.26	11.00
FLa-LLM(ours)	✓	**37.25**	37.13	**25.12**	**18.38**	**14.20**
Improvement		+0.81	-2.24	+0.19	+2.12	+3.20

Table 2: Experimental results on CSL-daily dataset. * denotes methods reproduced by (Yin et al., 2023). † denotes methods reproduced by (Zhou et al., 2021a). We bold the highest scores in the gloss-based setting and gloss-free setting. Improvement represents comparisons with the previous best gloss-free result.

Method	Gloss-Free	Rouge-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
TF-H2S (Alvarez et al.)	✓	-	17.40	7.69	3.97	2.21
SLT-IV (Tarrés et al., 2023)	✓	-	**34.01**	**19.30**	12.18	8.03
GloFE-VN (Lin et al., 2023)	✓	12.61	14.94	7.27	3.93	2.24
FLa-LLM(ours)	✓	**27.81**	29.81	18.99	**13.27**	**9.66**
Improvement		+15.20	-4.20	-0.31	+1.09	+1.63

Table 3: Experimental results on How2Sign dataset. We bold the highest scores. Improvement represents comparisons with the previous best gloss-free result.

Analysis



Dominance of LLM :

- The grad norm and parameter norm can reflect which part of the training process is more active.
- The main update of the model lies in the LLM module.

Ablation

Factorized	R	B1	B2	B3	B4
×	32.52	31.96	21.96	16.32	12.90
✓	**45.27**	**46.29**	**35.33**	**28.03**	**23.09**

Table 4: Effect of the proposed factorized learning strategy. The first row represents end-to-end joint training of the visual encoder and LLM.

VIS	LFS	R	B1	B2	B3	B4
×	✓	17.33	17.64	10.51	7.37	5.62
✓	×	38.67	39.09	28.20	21.83	17.69
✓	✓	**45.27**	**46.29**	**35.33**	**28.03**	**23.09**

Table 5: Effect of each stage. VIS represents the visual initialing stage and LFS means the LLM fine-tuning stage. The first row represents freezing the vision backbone and fine-tuning the other modules.

- **Factorized learning strategy substantially outperforms the end-to-end training.**
- **Based on sufficient initialization of the visual encoder, we successfully take advantage of the LLM and yield better results.**