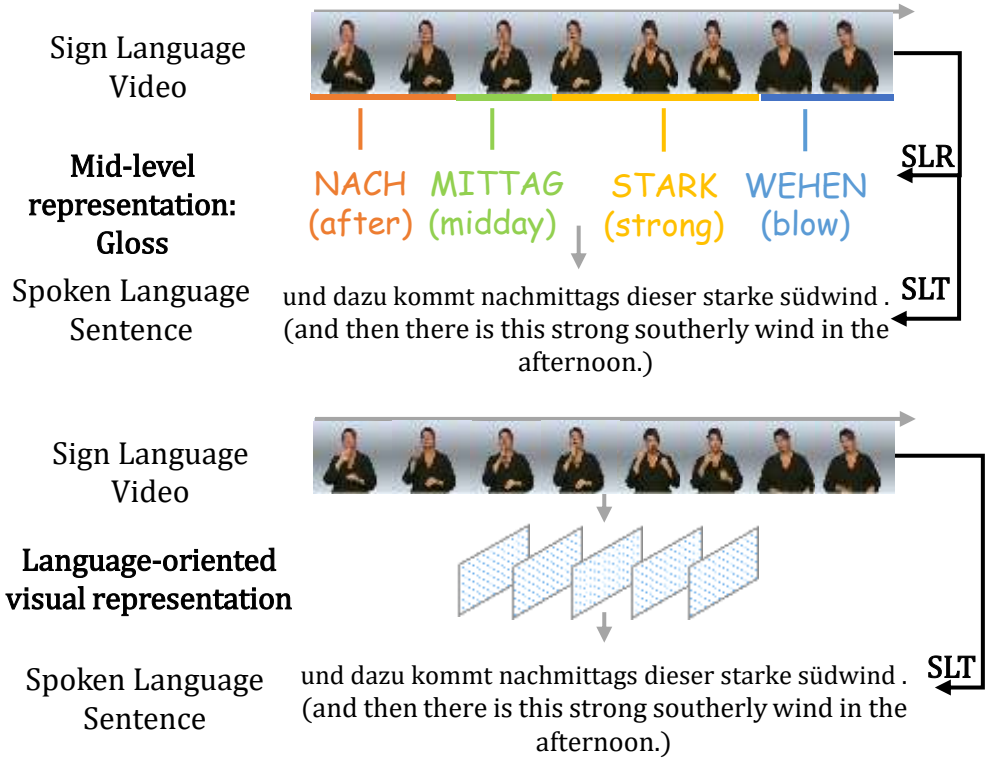
Factorized Learning Assisted with Large Language Model for Gloss-free Sign Language Translation

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

* Corresponding Author

Introduction

■ Gloss-free SLT



Gloss:

- **The transcription of sign languages.**
- **Every sign has a unique identifier**

Weakness of gloss-based:

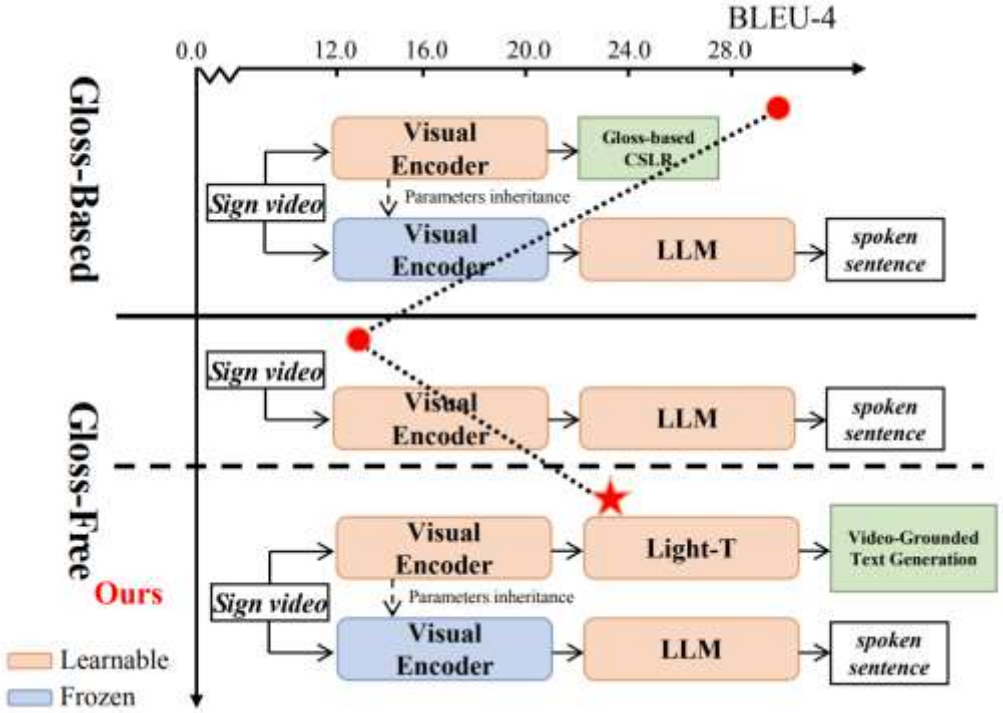
- **labor-intensive**
- **information bottleneck**

Gloss-free SLT

[1]Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, Du Zhang. Gloss-free Sign Language Translation: Improving from Visual-Language Pretraining. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)}, 2023.

Introduction

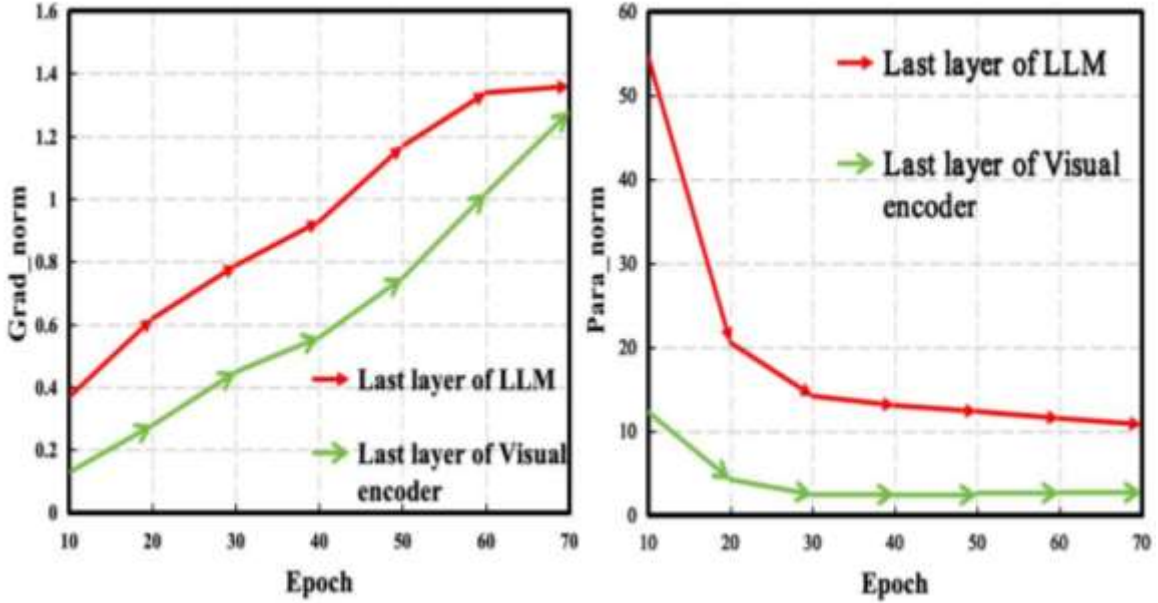
■ LLM for SLT



Contributions:

- **We analyze the reason why directly training the visual encoder and LLM failed in gloss-free SLT.**
- **We propose FLa-LLM to overcome above problem.**
- **Our approach greatly boosts the performance of the gloss-free SLT in three popular datasets.**

The Dominance of LLM in SLT

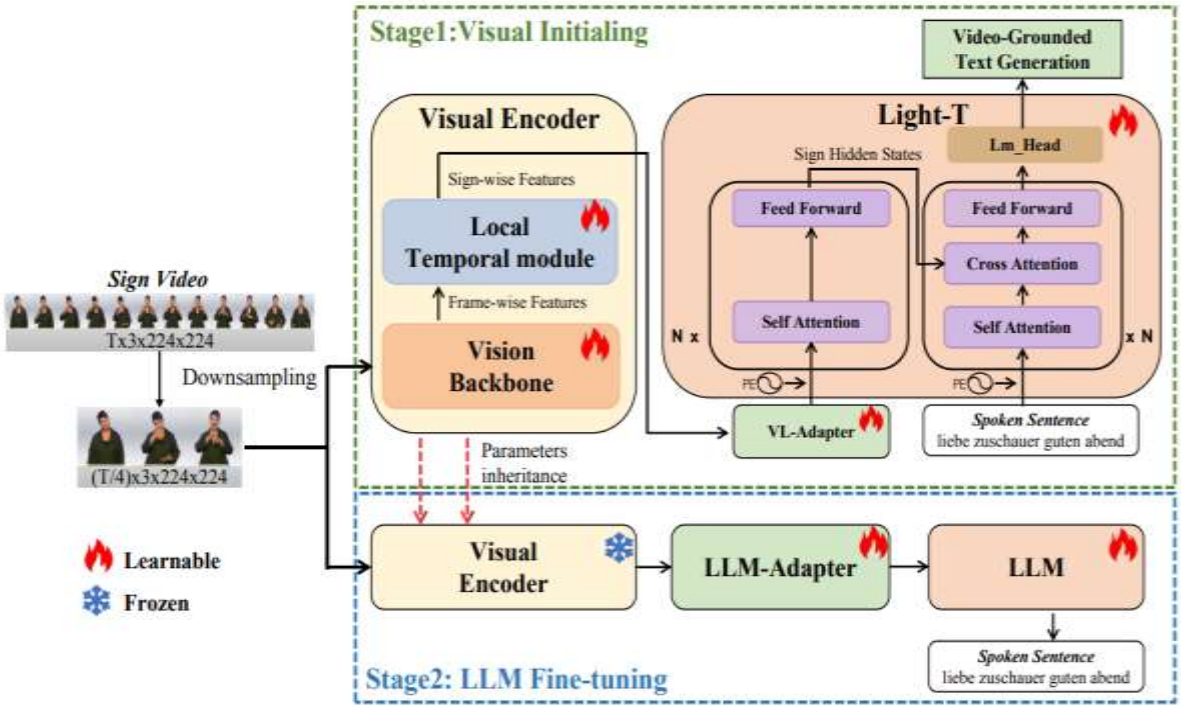


Dominance of LLM :

- **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.**

Method: FLa-LLM

■ Visual Initialing



Visual Initialing:

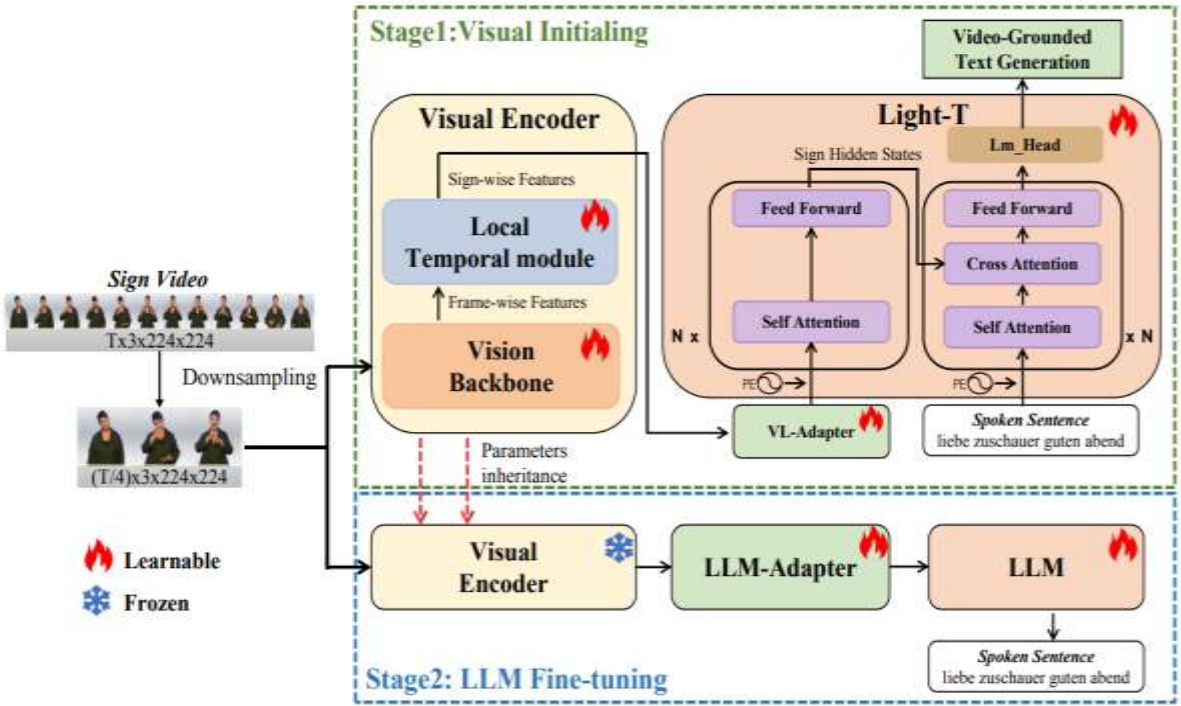
- **We construct a visual encoder followed by a lightweight translation model (Light-T) to perform visual initialing by a video-grounded text generation task.**

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

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

Method: FLa-LLM

■ LLM Fine-tuning



LLM Fine-tuning:

- **Freeze the Visual Encoder.**
- **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

Datasets

- **PHOENIX-2014T: German Sign Language (DGS), 8K.**
- **CSL-Daily: Chinese Sign Language (CSL), 20K.**
- **How2Sign: American Sign Language (ASL), 35K.**

Protocol.

- **Gloss-free SLT: a direct translation from sign language videos to the corresponding spoken sentences without gloss assistance through the entire framework.**

Evaluation Metrics.

- **ROUGE-L**
- **BLEU**

Experiments

■ Results on PHOENIX-2014T

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.

Experiments

■ Results on CSL-Daily

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.

Experiments

■ Results on How2Sign

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.

Experiments

■ Ablation on Factorized Learning

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.**

Experiments

■ Ablation on Visual Initializing

Rate	Time	R	B1	B2	B3	B4
100%	17.90h	44.55	45.68	35.01	27.82	22.96
50%	9.85h	44.60	46.22	35.12	27.90	23.04
25%	4.75h	**45.27**	**46.29**	**35.33**	**28.03**	**23.09**
12.5%	3.55h	40.77	42.42	31.62	24.68	20.02

Table 6: Effect of downsampling rate. The second column represents the time required to complete the visual initialing stage.

Size	Settings	Params	B4
Tiny	(1,4,256,1024)	3.61M	22.52
Small	(2,4,512,2048)	18.25M	22.36
Base	(3,8,512,2048)	25.61M	**23.09**
Large	(4,8,1024,4096)	124.66M	22.49

Table 7: Effect of translation network scale. The (1,4,256,1024) in the second column represents that the transformer has 1 hidden layer, 4 attention heads, a hidden size of 256, and a feed-forward dim of 2048. Params represents the number of model parameters.

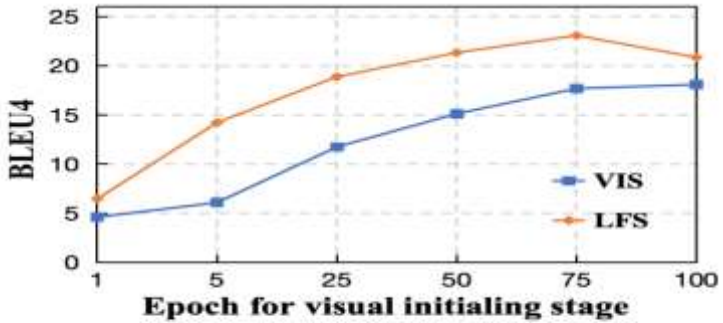


Figure 4: Effect of initialing time. VIS represents the visual initialing stage and LFS represents the LLM fine-tuning stage.

- **A sampling rate of 25% to ensure model performance and save training time.**
- **The transformer scale has little impact on final performance.**
- **We choose the model with epoch75 for fine-tuning.**

Experiments

■ Ablation on LLM Fine-tuning

Features	R	B1	B2	B3	B4
Frame-wise	41.59	42.54	31.72	24.82	20.30
Sign-wise	**45.27**	**46.29**	**35.33**	**28.03**	**23.09**
Hidden states	45.16	45.41	34.74	27.47	22.60

Table 8: Effect of different input features of LLM. The different features are shown in Figure 3.

VB	TM	R	B1	B2	B3	B4
×	×	40.72	40.74	29.79	22.65	17.86
✓	×	39.86	41.64	31.48	24.86	20.40
✓	✓	**45.27**	**46.29**	**35.33**	**28.03**	**23.09**

Table 9: Effect of freezing different parts of the visual encoder. VB means visual backbone. TM represents the local temporal module. ✓ means freezing the module while × means no freezing.

LLM	R	B1	B2	B3	B4
MBart w/o pre	40.18	37.18	26.99	20.54	16.19
MT5-Base w/o pre	22.71	18.02	12.21	9.17	7.39
MBart w/ pre	**45.27**	**46.29**	**35.33**	**28.03**	**23.09**
MT5-Base w/ pre	41.06	41.96	31.20	24.24	19.71

Table 10: Effect of different LLMs. W/o, w/, and pre means without, with, and pretraining, respectively.

- **The best result is obtained by using sign-wise features as input for LLM fine-tuning.**
- **The visual encoder already has a sufficient visual representation of sign language after the visual initializing stage.**
- **Pre-training on large-scale corpus can significantly improve the performance.**

Conclusion and Limition

Conclusion

- **We analyze the reason why directly training the visual encoder and LLM failed in gloss-free SLT.**
- **We propose FLa-LLM to overcome above problem.**
- **Our approach greatly boosts the performance of the gloss-free SLT in three popular datasets.**

Limitations

- **Two stage training is more cumbersome compared to end-to-end training.**
- **We fine-tune all parameters of the large language model which limits the scale of our LLM.**

ICCV23
PARIS

Thanks