JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Highlight Paper

Poster Number: 242

Tag: THU-PM-242

CVT-SLR: Contrastive Visual-Textual Transformation for Sign Language Recognition with Variational Alignment

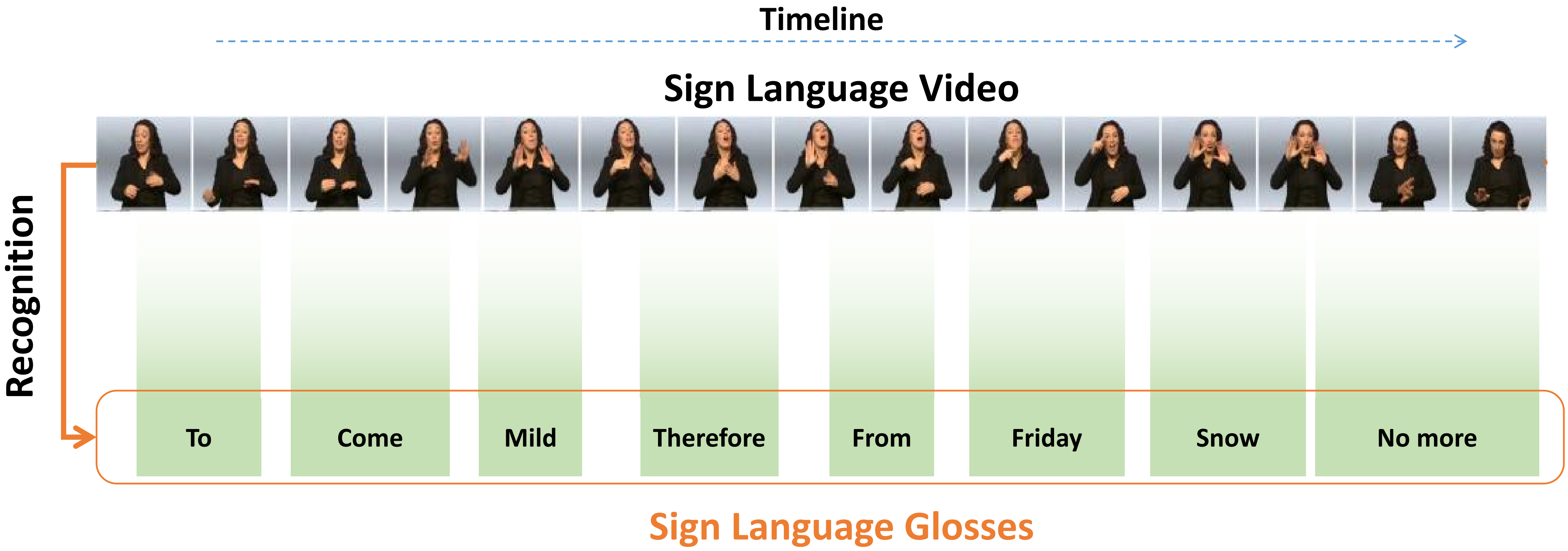
Jiangbin Zheng[1], Yile Wang[1,2], Cheng Tan[1], Siyuan Li[1], Ge Wang[1],
Jun Xia[1], Yidong Chen[3], Stan Z. Li[1,*]

[1]AI Lab, Research Center for Industries of the Future, Westlake University
[2]Institute for AI Industry Research (AIR), Tsinghua University
[3]School of Informatics, Xiamen University

Github：https://github.com/binbinjiang/CVT-SLR

西湖大學
WESTLAKE UNIVERSITY

Sign Language Recognition (SLR): Convert Sign Languge into Glosses



Timeline

Sign Language Video

Recognition

To

Come

Mild

Therefore

From

Friday

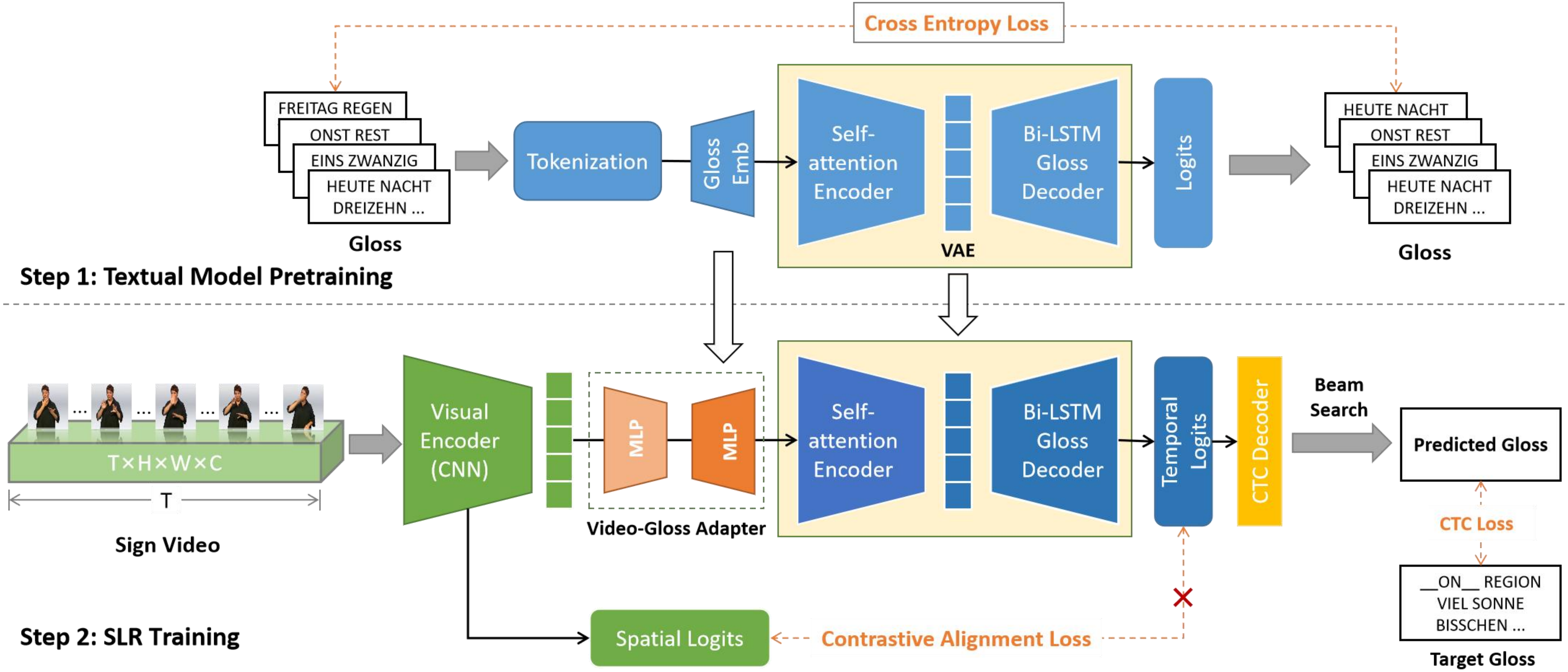
Snow

No more

Sign Language Glosses

SLR is a weakly supervised task due to lack of large-scale available sign datasets

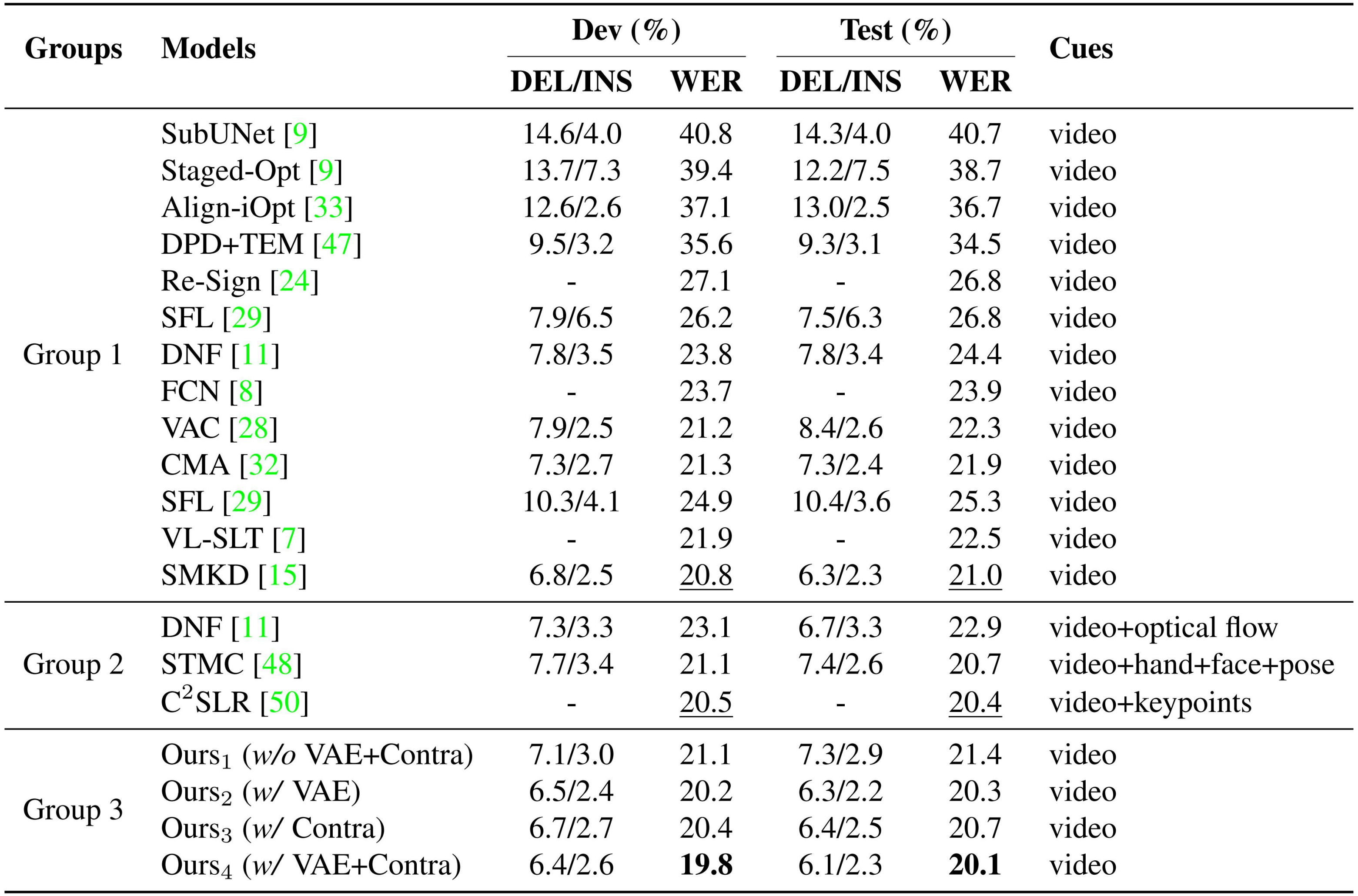
CVT-SLR Pipeline



Step 1: Textual Model Pretraining

Step 2: SLR Training

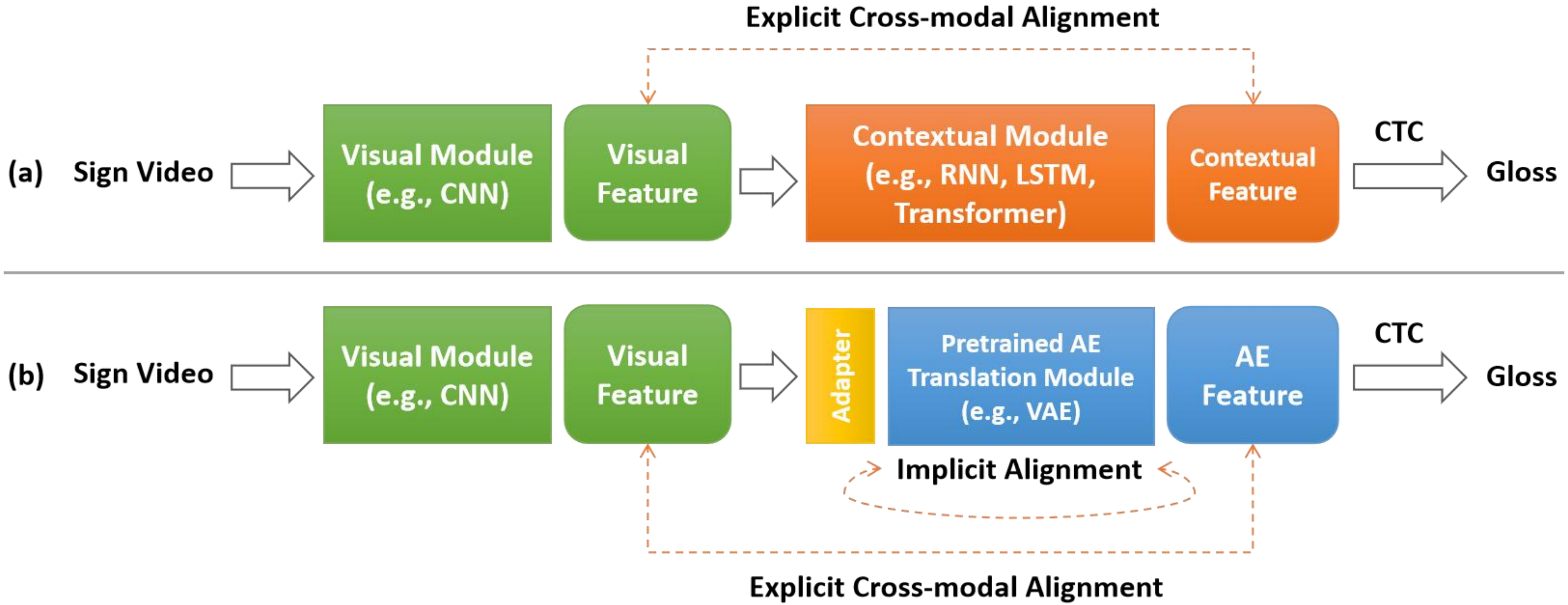
Main Results



Groups	Models	Dev (%)		Test (%)		Cues
		DEL/INS	WER	DEL/INS	WER	
Group 1	SubUNet [9]	14.6/4.0	40.8	14.3/4.0	40.7	video
	Staged-Opt [9]	13.7/7.3	39.4	12.2/7.5	38.7	video
	Align-iOpt [33]	12.6/2.6	37.1	13.0/2.5	36.7	video
	DPD+TEM [47]	9.5/3.2	35.6	9.3/3.1	34.5	video
	Re-Sign [24]	-	27.1	-	26.8	video
	SFL [29]	7.9/6.5	26.2	7.5/6.3	26.8	video
	DNF [11]	7.8/3.5	23.8	7.8/3.4	24.4	video
	FCN [8]	-	23.7	-	23.9	video
	VAC [28]	7.9/2.5	21.2	8.4/2.6	22.3	video
	CMA [32]	7.3/2.7	21.3	7.3/2.4	21.9	video
	SFL [29]	10.3/4.1	24.9	10.4/3.6	25.3	video
	VL-SLT [7]	-	21.9	-	22.5	video
	SMKD [15]	6.8/2.5	20.8	6.3/2.3	21.0	video
Group 2	DNF [11]	7.3/3.3	23.1	6.7/3.3	22.9	video+optical flow
	STMC [48]	7.7/3.4	21.1	7.4/2.6	20.7	video+hand+face+pose
	C^{2}SLR [50]	-	20.5	-	20.4	video+keypoints
Group 3	Ours$_1$ (*w/o* VAE+Contra)	7.1/3.0	21.1	7.3/2.9	21.4	video
	Ours$_2$ (*w/* VAE)	6.5/2.4	20.2	6.3/2.2	20.3	video
	Ours$_3$ (*w/* Contra)	6.7/2.7	20.4	6.4/2.5	20.7	video
	Ours$_4$ (*w/* VAE+Contra)	6.4/2.6	**19.8**	6.1/2.3	**20.1**	video

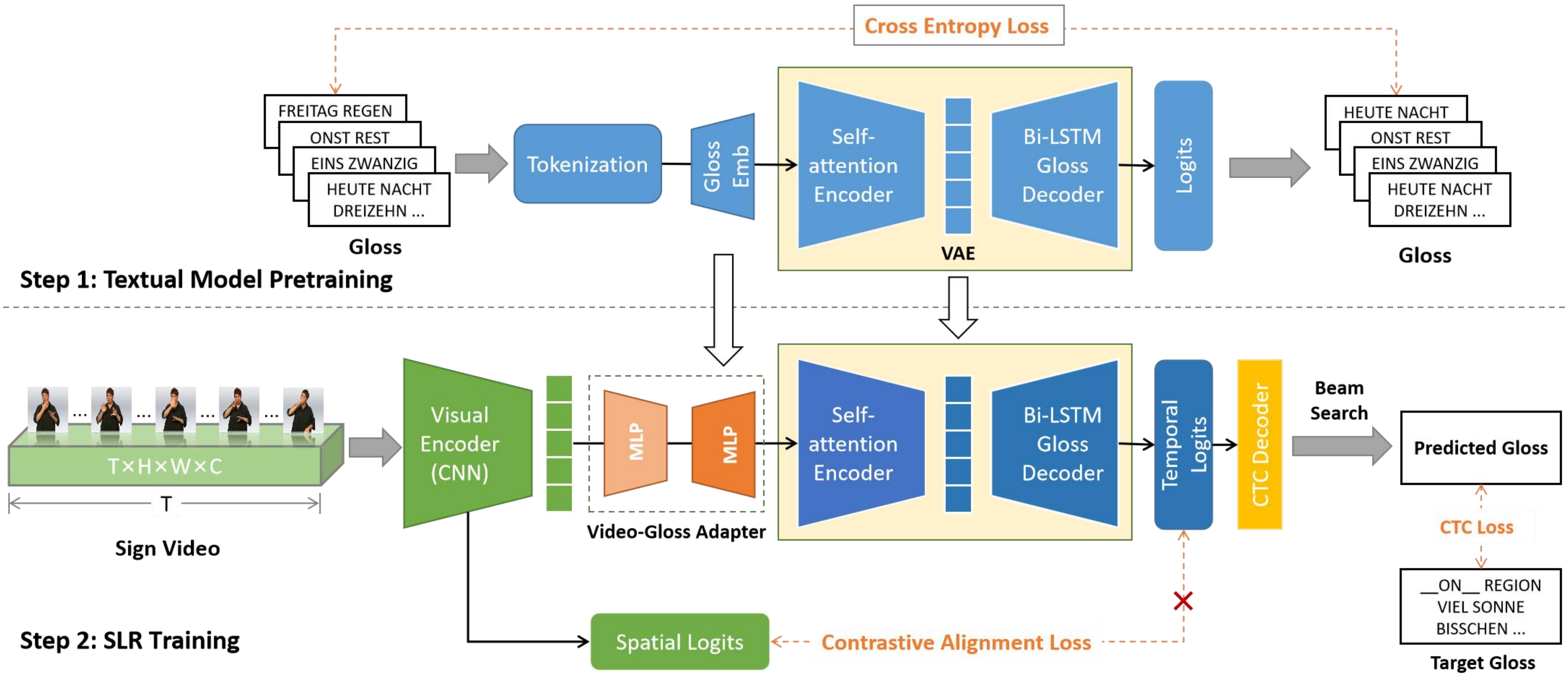
Performance comparison on PHOENIX-2014 dataset.

Introduction


Explicit Cross-modal Alignment
(a)
Sign Video
Visual Module (e.g., CNN)
Visual Feature
Contextual Module (e.g., RNN, LSTM, Transformer)
Contextual Feature
CTC
Gloss
(b)
Sign Video
Visual Module (e.g., CNN)
Visual Feature
Adapter
Pretrained AE Translation Module (e.g., VAE)
AE Feature
CTC
Gloss
Implicit Alignment
Explicit Cross-modal Alignment

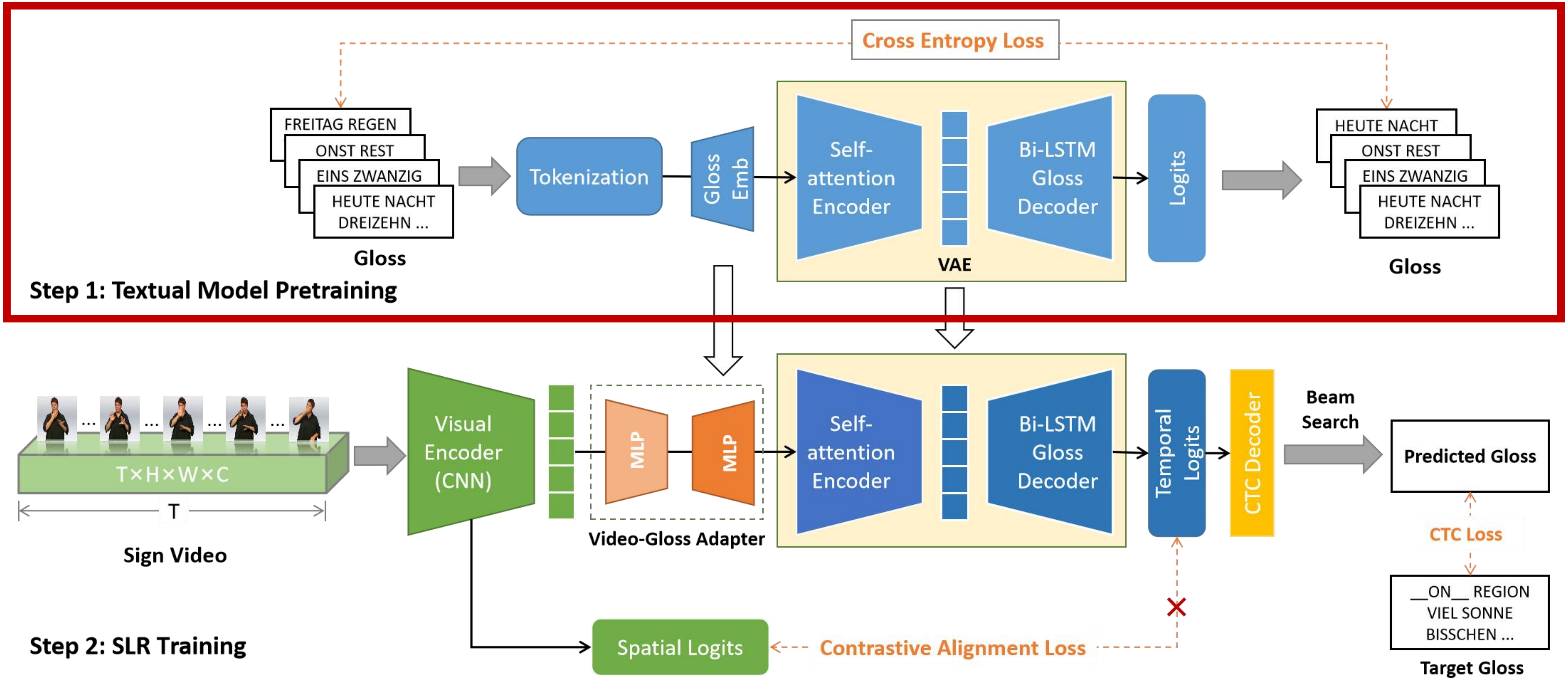
(a) An advanced single-cue SLR framework with explicit cross-modal alignment;
(b) Our proposed single-cue SLR framework with explicit cross-modal alignment and implicit autoencoder alignment.

Methods

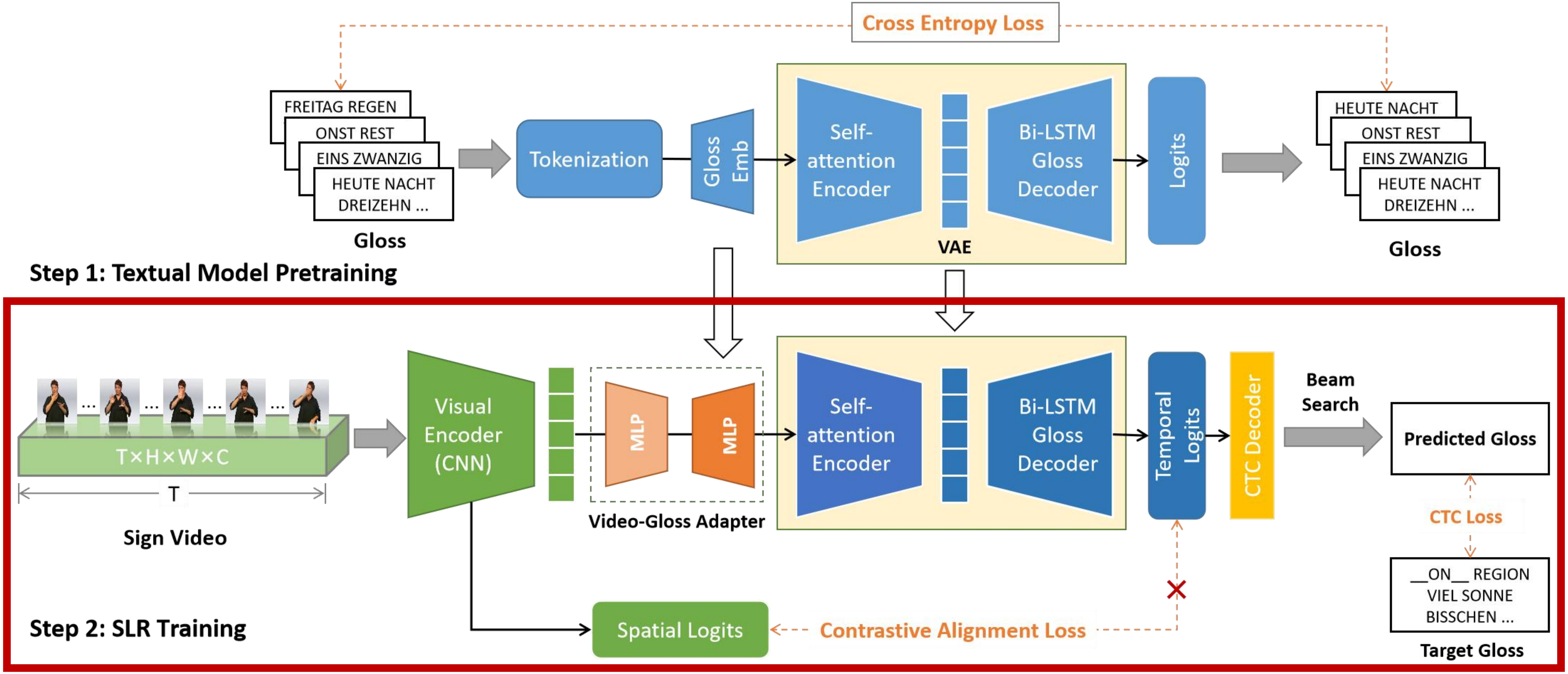


Cross Entropy Loss
FREITAG REGEN
ONST REST
EINS ZWANZIG
HEUTE NACHT DREIZEHN ...
Gloss
Tokenization
Gloss Emb
Self-attention Encoder
VAE
Bi-LSTM Gloss Decoder
Logits
HEUTE NACHT
ONST REST
EINS ZWANZIG
HEUTE NACHT DREIZEHN ...
Gloss
Step 1: Textual Model Pretraining
T×H×W×C
T
Sign Video
Visual Encoder (CNN)
MLP
MLP
Video-Gloss Adapter
Self-attention Encoder
Bi-LSTM Gloss Decoder
Temporal Logits
CTC Decoder
Beam Search
Predicted Gloss
CTC Loss
__ON__ REGION VIEL SONNE BISSCHEN ...
Target Gloss
Spatial Logits
Contrastive Alignment Loss
Step 2: SLR Training

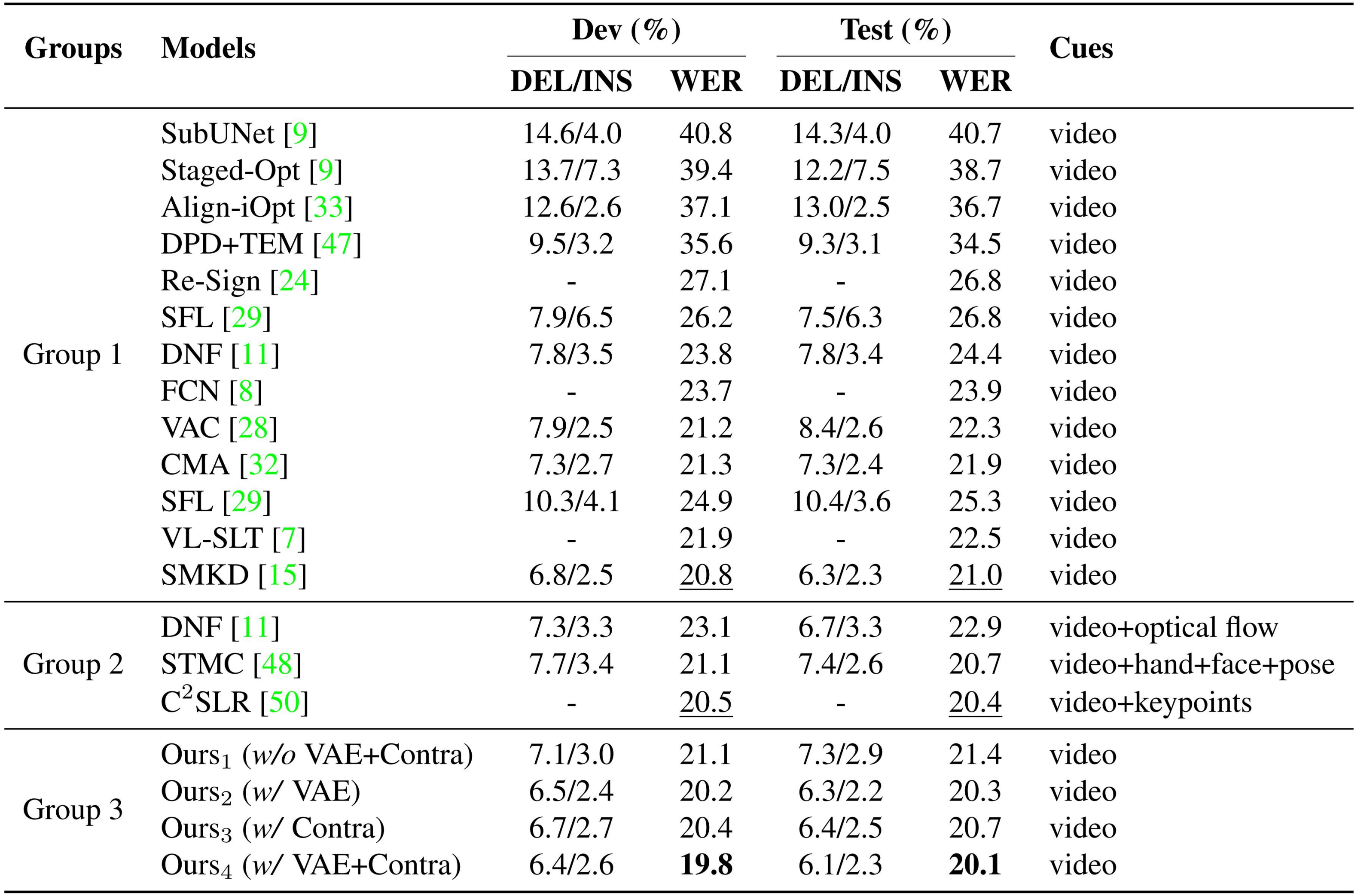
Methods


Cross Entropy Loss
FREITAG REGEN
ONST REST
EINS ZWANZIG
HEUTE NACHT
DREIZEHN ...
Gloss
Tokenization
Gloss Emb
Self-attention Encoder
VAE
Bi-LSTM Gloss Decoder
Logits
HEUTE NACHT
ONST REST
EINS ZWANZIG
HEUTE NACHT
DREIZEHN ...
Gloss
Step 1: Textual Model Pretraining
T×H×W×C
T
Sign Video
Visual Encoder (CNN)
MLP
MLP
Video-Gloss Adapter
Self-attention Encoder
Bi-LSTM Gloss Decoder
Temporal Logits
CTC Decoder
Beam Search
Predicted Gloss
CTC Loss
__ON__ REGION
VIEL SONNE
BISSCHEN ...
Target Gloss
Spatial Logits
Contrastive Alignment Loss
Step 2: SLR Training

Methods


Cross Entropy Loss
FREITAG REGEN
ONST REST
EINS ZWANZIG
HEUTE NACHT
DREIZEHN ...
Gloss
Tokenization
Gloss Emb
Self-attention Encoder
VAE
Bi-LSTM Gloss Decoder
Logits
HEUTE NACHT
ONST REST
EINS ZWANZIG
HEUTE NACHT
DREIZEHN ...
Gloss
Step 1: Textual Model Pretraining
T×H×W×C
T
Sign Video
Visual Encoder (CNN)
MLP
MLP
Video-Gloss Adapter
Self-attention Encoder
Bi-LSTM Gloss Decoder
Temporal Logits
CTC Decoder
Beam Search
Predicted Gloss
CTC Loss
__ON__ REGION
VIEL SONNE
BISSCHEN ...
Target Gloss
Spatial Logits
Contrastive Alignment Loss
Step 2: SLR Training

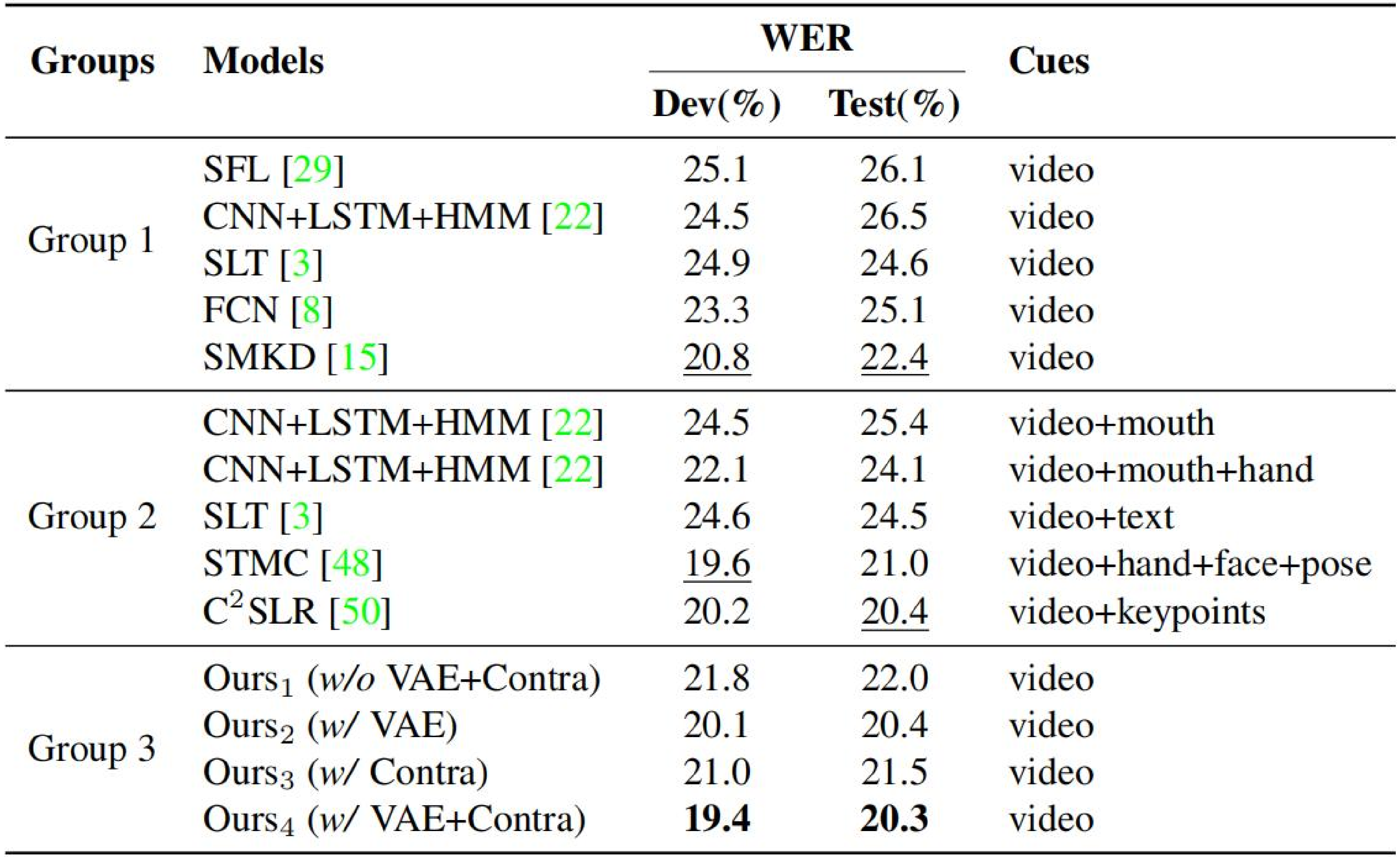
Experiments



Groups	**Models**	**Dev (%)**		**Test (%)**		**Cues**
		DEL/INS	**WER**	**DEL/INS**	**WER**	
Group 1	SubUNet [9]	14.6/4.0	40.8	14.3/4.0	40.7	video
	Staged-Opt [9]	13.7/7.3	39.4	12.2/7.5	38.7	video
	Align-iOpt [33]	12.6/2.6	37.1	13.0/2.5	36.7	video
	DPD+TEM [47]	9.5/3.2	35.6	9.3/3.1	34.5	video
	Re-Sign [24]	-	27.1	-	26.8	video
	SFL [29]	7.9/6.5	26.2	7.5/6.3	26.8	video
	DNF [11]	7.8/3.5	23.8	7.8/3.4	24.4	video
	FCN [8]	-	23.7	-	23.9	video
	VAC [28]	7.9/2.5	21.2	8.4/2.6	22.3	video
	CMA [32]	7.3/2.7	21.3	7.3/2.4	21.9	video
	SFL [29]	10.3/4.1	24.9	10.4/3.6	25.3	video
	VL-SLT [7]	-	21.9	-	22.5	video
	SMKD [15]	6.8/2.5	<u>20.8</u>	6.3/2.3	<u>21.0</u>	video
Group 2	DNF [11]	7.3/3.3	23.1	6.7/3.3	22.9	video+optical flow
	STMC [48]	7.7/3.4	21.1	7.4/2.6	20.7	video+hand+face+pose
	C^{2}SLR [50]	-	<u>20.5</u>	-	<u>20.4</u>	video+keypoints
Group 3	Ours$_1$ (*w/o* VAE+Contra)	7.1/3.0	21.1	7.3/2.9	21.4	video
	Ours$_2$ (*w/* VAE)	6.5/2.4	20.2	6.3/2.2	20.3	video
	Ours$_3$ (*w/* Contra)	6.7/2.7	20.4	6.4/2.5	20.7	video
	Ours$_4$ (*w/* VAE+Contra)	6.4/2.6	**19.8**	6.1/2.3	**20.1**	video

Table 1. Performance comparison (%) on PHOENIX-14 dataset. DEL/INS: deletion error and insertion error. The best results and SOTA baseline for each group are marked as bold and underlined.

Experiments

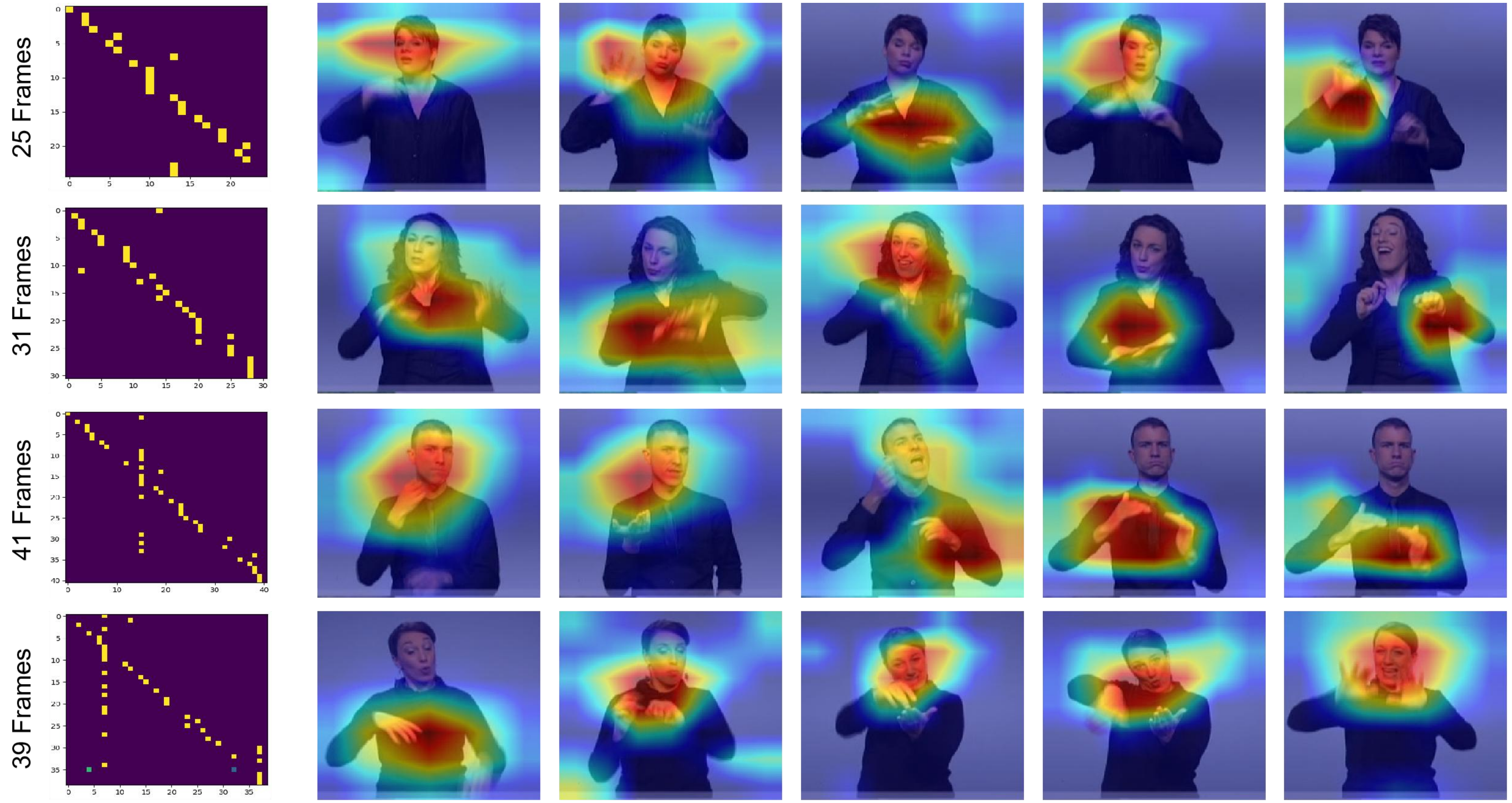


Groups	Models	WER Dev(%)	WER Test(%)	Cues
Group 1	SFL [29]	25.1	26.1	video
	CNN+LSTM+HMM [22]	24.5	26.5	video
	SLT [3]	24.9	24.6	video
	FCN [8]	23.3	25.1	video
	SMKD [15]	20.8	22.4	video
Group 2	CNN+LSTM+HMM [22]	24.5	25.4	video+mouth
	CNN+LSTM+HMM [22]	22.1	24.1	video+mouth+hand
	SLT [3]	24.6	24.5	video+text
	STMC [48]	19.6	21.0	video+hand+face+pose
	C^2SLR [50]	20.2	20.4	video+keypoints
Group 3	Ours$_1$ (*w/o* VAE+Contra)	21.8	22.0	video
	Ours$_2$ (*w/* VAE)	20.1	20.4	video
	Ours$_3$ (*w/* Contra)	21.0	21.5	video
	Ours$_4$ (*w/* VAE+Contra)	**19.4**	**20.3**	video

Table 2. Performance comparison (%) on PHOENIX-14T dataset. The best results and SOTA baseline for each group are marked as bold and underlined, respectively.

Experiments

Visualization



Example 1：

Infer: __ON__ FREITAG REGEN WECHSELHAFT loc-NORDWEST KOENNEN GEWITTER __OFF__

Ref: __ON__ FREITAG REGEN WECHSELHAFT NORDWEST KOENNEN GEWITTER __OFF__

Example 2：

Infer: DONNERSTAG cl-KOMMEN REGEN KOENNEN GLATT

Ref: DONNERSTAG KOMMEN REGEN KOENNEN GLATT

Example 3：

Infer: AUCH BISSCHEN SONNE MORGEN MEISTENS REGEN TAG ANDERE cl-KOMMEN

Ref: AUCH BISSCHEN SONNE MORGEN MEISTENS REGEN TAG IM-VERLAUF KOMMEN

Example 4：

Infer: __ON__ HEUTE NACHT __EMOTION__ NOCH STURM MITTE SUEDRAUM DAZU SCHNEE REGEN

Ref: __ON__ HEUTE NACHT NOCH STURM MITTE SUED DAZU SCHNEE REGEN

Four examples with cross-modal alignment matrices (left), saliency maps (middle), and generated glosses (right) on the PHOENIX-14 test set.

Conclusion

- A novel visual-textual transformation-based SLR framework is proposed
- New alignment methods are proposed for cross-modal consistency constraints.
- The proposed single-cue SLR framework not only outperforms existing baselines by a large margin.
- The source codes and models are available at: *https://github.com/binbinjiang/CVT-SLR*

Highlight Paper

CVT-SLR: Contrastive Visual-Textual Transformation for Sign Language Recognition with Variational Alignment

THANKS

西湖大學
WESTLAKE UNIVERSITY