GLOSS SEMANTIC-ENHANCED NETWORK WITH ONLINE BACK-TRANSLATION FOR SIGN LANGUAGE PRODUCTION

Shengeng Tang, Richang Hong, Dan Guo, Meng Wang
Hefei University of Technology, Hefei, China
Address: tsg1995@mail.hfut.edu.cn
Paper ID: mmfp0365

ACM
Multimedia 2022
Lisbon, Portugal | 10-14 October

Backgrounds and Proposed Idea

Sign Language Production (SLP) aims to generate the visual appearance of sign language according to the spoken language, in which a key procedure is to translate sign Gloss to Pose (G2P). Existing G2P methods mainly focus on regression prediction of posture coordinates, namely closely fitting the ground truth.

In this work, we provide a new viewpoint: a Gloss semantic-Enhanced Network is proposed with Online Back-Translation (GEN-OBT) for G2P. GEN-OBT consists of a gloss encoder, a pose decoder, and an online reverse gloss decoder. Specifically, we design a learnable gloss token without any prior knowledge, to explore the global contextual dependency of the entire gloss sequence. Furthermore, we design a CTC-based reverse decoder to convert the generated poses backward into glosses, which guarantees the semantic consistency during the processes of gloss-to-pose and pose-to-gloss.

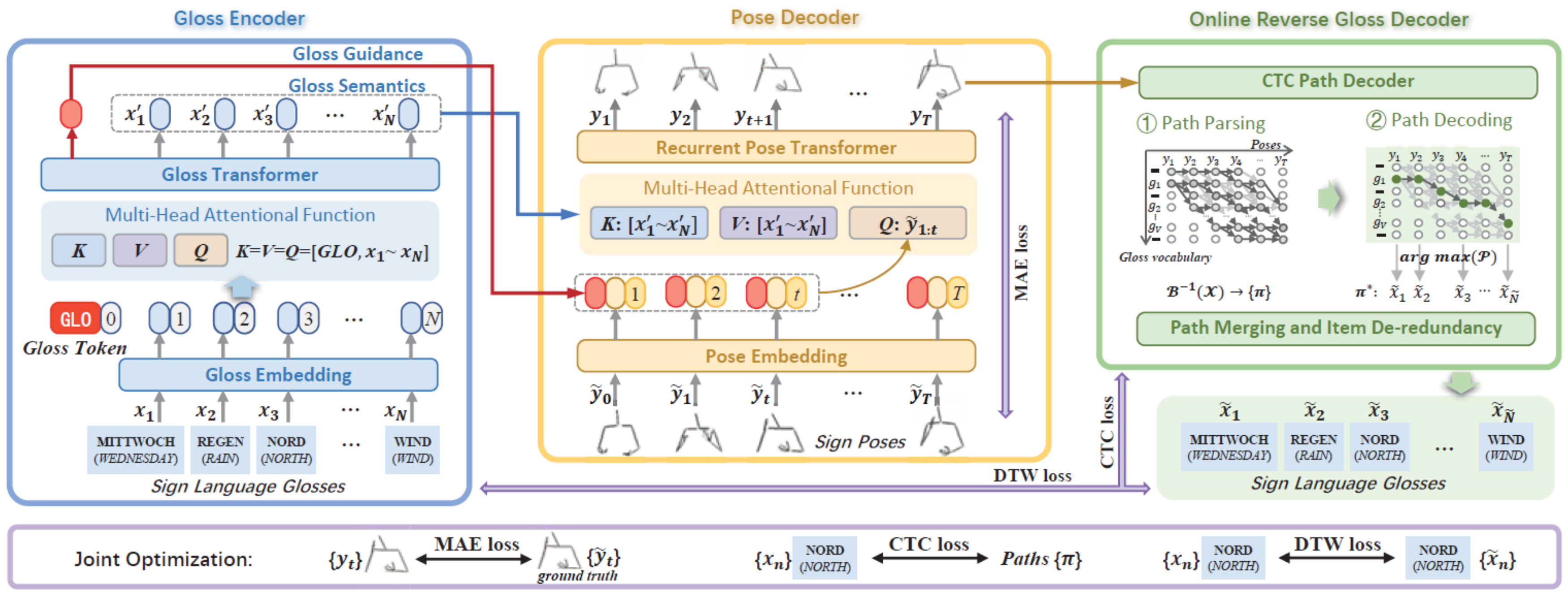

Gloss Encoder
Gloss Guidance
Gloss Semantics
x'_1 x'_2 x'_3 ... x'_N
Gloss Transformer
Multi-Head Attentional Function
K V Q $K=V=Q=[GLO, x_1 \sim x_N]$
GLO 0 1 2 3 ... N
Gloss Token
Gloss Embedding
x_1 x_2 x_3 ... x_N
MITTWOCH (WEDNESDAY) REGEN (RAIN) NORD (NORTH) ... WIND (WIND)
Sign Language Glosses
Pose Decoder
y_1 y_2 y_{t+1} ... y_T
Recurrent Pose Transformer
Multi-Head Attentional Function
K: $[x'_1 \sim x'_N]$ V: $[x'_1 \sim x'_N]$ Q: $\tilde{y}_{1:t}$
1 2 t ... T
Pose Embedding
$\tilde{y}_0$ $\tilde{y}_1$ $\tilde{y}_t$ ... $\tilde{y}_T$
Sign Poses
MAE loss
DTW loss
CTC loss
Online Reverse Gloss Decoder
CTC Path Decoder
① Path Parsing
② Path Decoding
Poses
Gloss vocabulary
$\arg\max(\mathcal{P})$
$\mathcal{B}^{-1}(\mathcal{X}) \rightarrow \{\pi\}$
π^*: $\tilde{x}_1$ $\tilde{x}_2$ $\tilde{x}_3$ ... $\tilde{x}_{\tilde{N}}$
Path Merging and Item De-redundancy
$\tilde{x}_1$ $\tilde{x}_2$ $\tilde{x}_3$ $\tilde{x}_{\tilde{N}}$
MITTWOCH (WEDNESDAY) REGEN (RAIN) NORD (NORTH) ... WIND (WIND)
Sign Language Glosses
Joint Optimization:
$\{y_t\}$ MAE loss $\{\tilde{y}_t\}$ ground truth
$\{x_n\}$ NORD (NORTH) CTC loss Paths $\{\pi\}$
$\{x_n\}$ NORD (NORTH) DTW loss NORD (NORTH) $\{\tilde{x}_n\}$

Fig.1: Overview of the proposed framework - GEN-OBT.

Framework of The Proposed Method

Given a sign gloss sentence with N glosses $\mathcal{X} = \{x_1, x_2, \cdots, x_N\}$, SLP is required to generate a pose sequence $\mathcal{Y} = \{y_1, y_2, \cdots, y_T\}$, where T is the number of generated poses.

The proposed GEN-OBT framework includes three modules: a *Gloss Encoder*, a *Pose Decoder*, and an online reverse *Gloss Decoder*. We apply the transformer architecture as the network backbone. In the *Gloss Encoder*, after gloss embedding, we use a learnable token, named "gloss token", to capture the global semantics of the gloss sequence. Then, in the *Pose Decoder*, we add up gloss token to the pose embedding vectors and take it as *Query*, and further take the gloss embedding sequence as both *Key* and *Value* in a recurrent transformer. In other words, we leverage previous poses and the entire gloss sequence to predict the next pose. In the reverse Gloss, we calculate the probability of each generated pose over the gloss vocabulary and decode the alignment paths of pose-to-gloss by Connectionist Temporal Classification (CTC) optimization.

As shown in *Fig.1*, loss $\mathcal{L}_{MAE}$ is proposed to constraint the coordinate consistency of generated poses and the ground-truth; loss $\mathcal{L}_{CTC}$ optimizes all the alignments of pose-to-gloss during back-translation; while loss $\mathcal{L}_{DTW}$ measures the matching score of the reproduced glosses with the original gloss sequences. These losses guarantee semantic preservation during pose generation and gloss back-translation.

Experimental Results

1) Visualization of Interactive Cross-modal Attention

We provide an example of cross-modal interaction between glosses and poses in the *Pose Decoder*. As shown in Fig.2, the highly responsive attention regions are distributed diagonally.

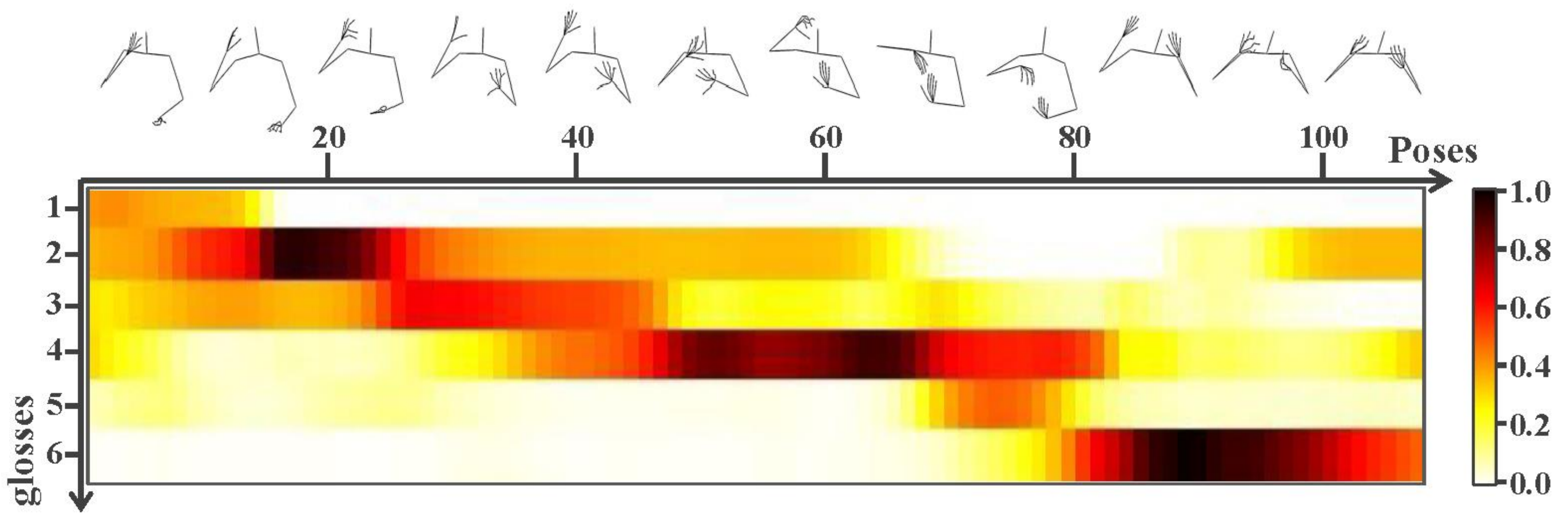
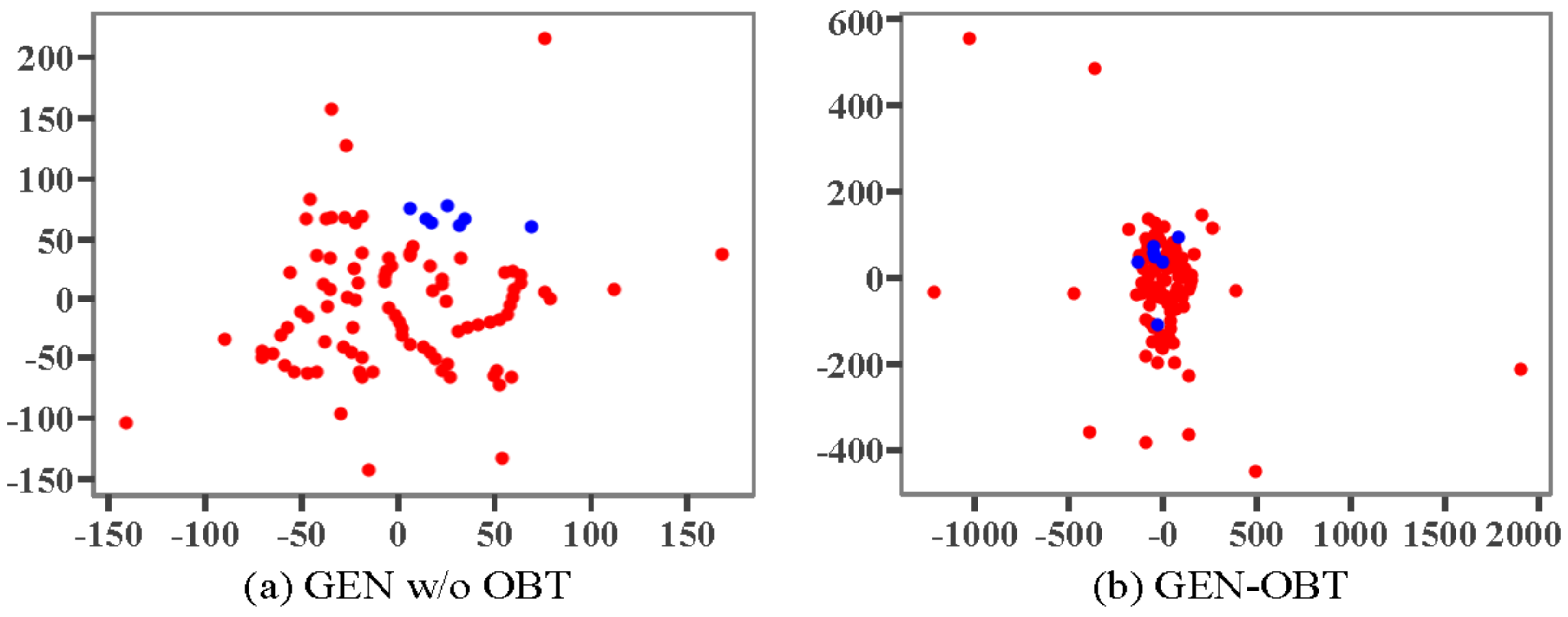


Fig.2: Visualization examples of the cross-modal interaction.

2) Visualization of Feature Distributions

We show feature distributions by t-SNE. Red and blue points mark the generated poses and original input glosses, respectively.



(a) GEN w/o OBT
(b) GEN-OBT

Fig.3: Visualization examples of the feature distributions.

3) Visualization Examples of Pose Sequences Generation

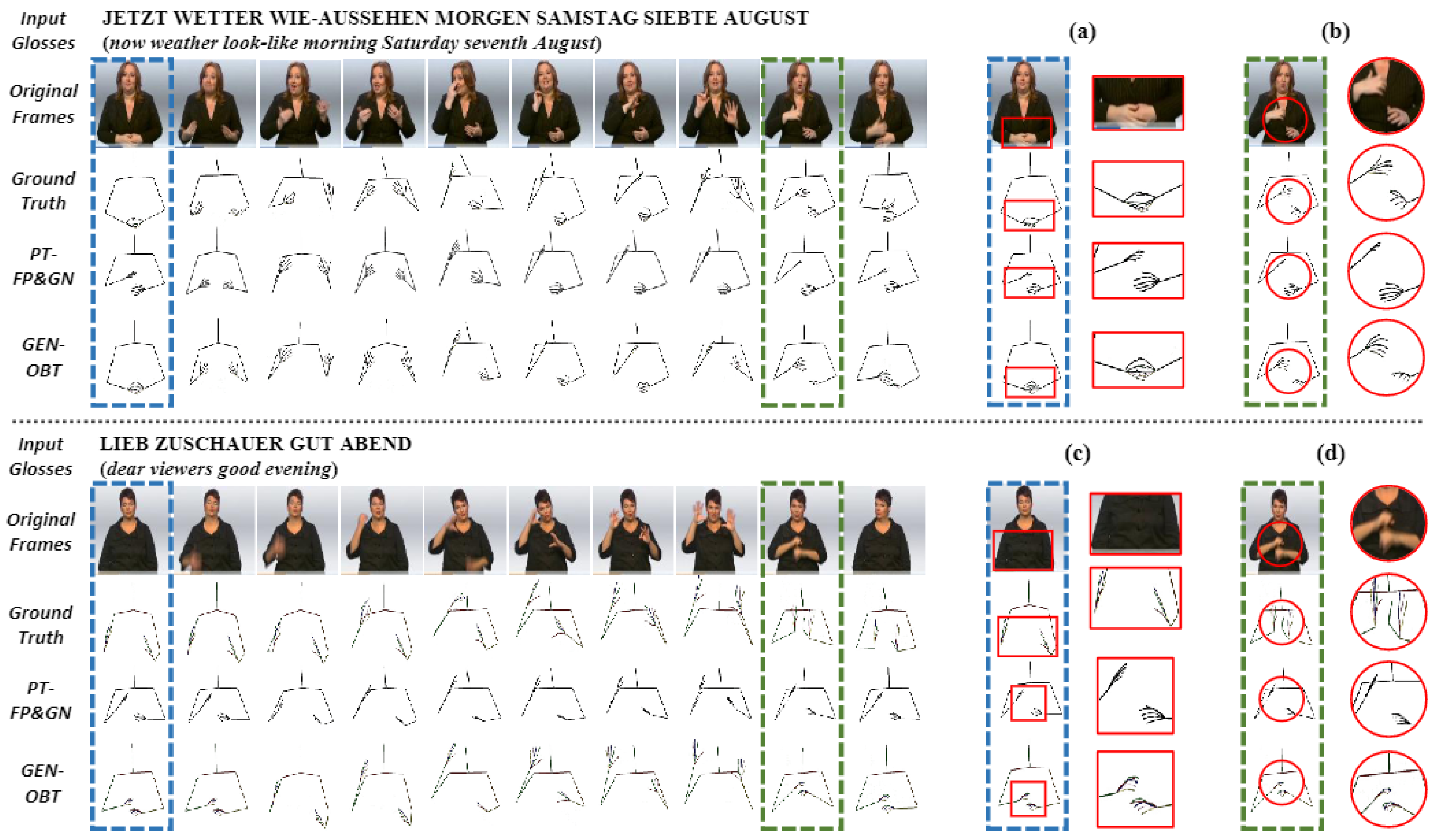


Fig.4: Visualization examples of the produced pose sequence.

4) Quantitative Evaluation of Produced Pose Sequences

We compare our GEN-OBT with state-of-the-arts. As shown in Tab.1, GEN-OBT performs prominent superiority over the others.

Methods	DEV						TEST					
	BLEU-1	BLEU-2	BLEU-3	ROUGE	WER↓	DTW-P↓	BLEU-1	BLEU-2	BLEU-3	ROUGE	WER↓	DTW-P↓
Ground Truth	29.77	20.21	15.16	29.60	74.17	0.00	29.76	20.12	14.93	28.98	71.94	0.00
PT-base[†]	9.53	3.45	1.62	8.61	98.53	29.33	9.47	3.37	1.47	8.88	98.36	28.48
PT-FP&GN[†]	12.51	6.50	4.76	11.87	96.85	11.75	13.35	7.29	5.33	13.17	96.50	11.54
NAT-AT	–	–	–	–	–	–	14.26	9.93	7.11	18.72	88.15	–
NAT-EA	–	–	–	–	–	–	15.12	10.45	7.99	19.43	82.01	–
DET	17.25	10.17	7.04	17.85	–	–	17.18	10.39	7.39	17.64	–	–
GEN-OBT	**24.92**	**15.72**	**11.20**	**25.21**	**82.36**	**10.37**	**23.08**	**14.91**	**10.48**	**23.49**	**81.78**	**10.07**

Tab.1: Quantitative results on PHOENIX14T dataset.