SignPR: A Progressive Vector-Quantized Diffusion Framework for Sign Language Production

Supplementary Material

A. Pose Length Prediction Module

To estimate the number of pose frames required for generation, we introduce a lightweight pose length prediction module. Given an input sentence c, we first use a pretrained BERT encoder to extract contextual text features. These features are further processed by a 2-layer Transformer encoder to capture long-range dependencies. The output is passed through a fully connected layer $\mathcal{FC}$ to predict the target sequence length $\hat{S}$:

$$\hat{S} = \mathcal{FC}(\text{Transformer}(\text{BERT}(c))) \tag{1}$$

The module is supervised using an L_1 loss $||*||_1$ between the predicted length $\hat{S}$ and the ground-truth length S:

$$\mathcal{L}_{\text{len}} = \|\hat{S} - S\|_1 \tag{2}$$

During inference, the predicted length $\hat{S}$ is used to determine the number of generated pose frames in the semantic and regional diffusion stage.

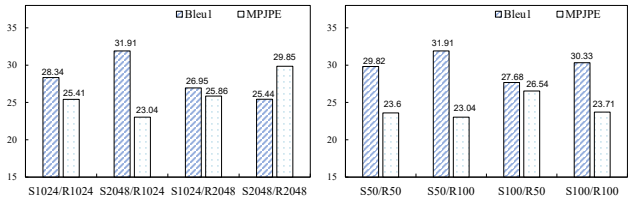
Performance of Pose Length Prediction. We evaluate the pose length prediction module on the test sets of all three datasets. As shown in Table 9, the mean absolute errors (MAE) are 6.42 frames on Phoenix14T, 9.15 frames on CSL-Daily, and 12.74 frames on USTC-CSL. To normalize the prediction error across sequences of varying lengths, we further report the Relative Error, computed as:

$$\text{Relative Error} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{y}_i - y_i|}{y_i}$$

where N denotes the number of test samples, y_i is the ground-truth length of the i-th sequence, and $\hat{y}_i$ is its predicted length. These low relative errors confirm the reliability of predicted lengths in guiding the generation process.

Dataset	MAE (frames)	Relative Error
Phoenix14T	6.42	5.89%
CSL-Daily	9.15	6.14%
USTC-CSL	12.74	4.06%

Table 9. MAE and Relative Error (%) of pose length prediction.



(a) Ablation study on codebook sizes in S-VQVAE (b) Ablation study on the diffusion steps

Figure 8. Ablation study on codebook sizes in S-VQVAE and diffusion steps.

B. More Experiments

Effect of Codebook Size Settings. As shown in Figure 8 (a), increasing the size of the semantic codebook from 1024 to 2048 while keeping regional fixed at 1024 consistently improves both BLEU1 and MPJPE, whereas enlarging the regional codebook to 2048 brings no further gains and even degrades performance. Therefore, we adopt S2048/R1024 as the default configuration.

Effect of Diffusion Steps. As shown in Figure 8 (b), increasing the number of inference steps in the regional diffusion module consistently improves performance, while changes in inference steps in semantic diffusion module have a worse impact. Using 50 diffusion steps in the semantic module and 100 in the regional module yields the best BLEU and ROUGE scores, *i.e.*, highlighting the regional stage benefits detailed pose generation.

C. User Study

To evaluate the quality of the generated videos, we conducted a user study involving 10 participants (5 signers and 5 non-signers). For each sample, participants were presented with one ground-truth video and three anonymized generated videos corresponding to PT, MoMP, and our SignPR model. The order of the generated videos was randomized for each sample to avoid position bias. Participants were asked to select the sample that best matched the ground-truth videos. In total, the study collected 200 votes on Phoenix14T, 200 on CSL-Daily, and 100 on USTC-CSL. All evaluations were conducted independently, and model identities were concealed throughout the process.

As shown in Table 10. Across all three datasets, SignPR consistently received the highest user preference. On Phoenix14T, SignPR achieved 73.5% of the votes (147 out of 200), significantly surpassing MoMP (17.5%) and PT

(9.0%). Consistent with the results on Phoenix14T, SignPR was favored by 68.5% of users on CSL-Daily. Even on USTC-CSL, with only 100 evaluated samples, SignPR remained dominant with 63.0% of the votes. These results confirm the effectiveness of SignPR in producing accurate and natural sign gestures.

D. More Qualitative Results

While the main paper provides qualitative results on Phoenix14T, here we present additional results on all three datasets: Phoenix14T, CSL-Daily, and USTC-CSL, as shown in Figure 9, 10, 11, 12, comparing ground truth (GT), SignPR, and a variant of SignPR without regional refinement. While the variant captures the overall structure, it often fails in specific frames, with errors in limb positioning and imprecise hand or facial expressions (highlighted by red and yellow boxes). In contrast, SignPR generates poses that are both semantically consistent and visually precise, correcting structural errors and refining regional details such as hand shapes and facial movements. These improvements demonstrate the effectiveness of our progressive diffusion framework in enhancing sign accuracy and expressiveness.

Method	Phoenix14T (200)		CSL-Daily (200)		USTC-CSL (100)	
	Votes	%	Votes	%	Votes	%
PT	18	9.0%	17	8.5%	7	7.0%
MoMP	35	17.5%	46	23.0%	30	30.0%
SignPR (Ours)	**147**	**73.5%**	**137**	**68.5%**	**63**	**63.0%**

Table 10. User preference results across three datasets. Each method's performance is reported as vote counts and selection percentage.

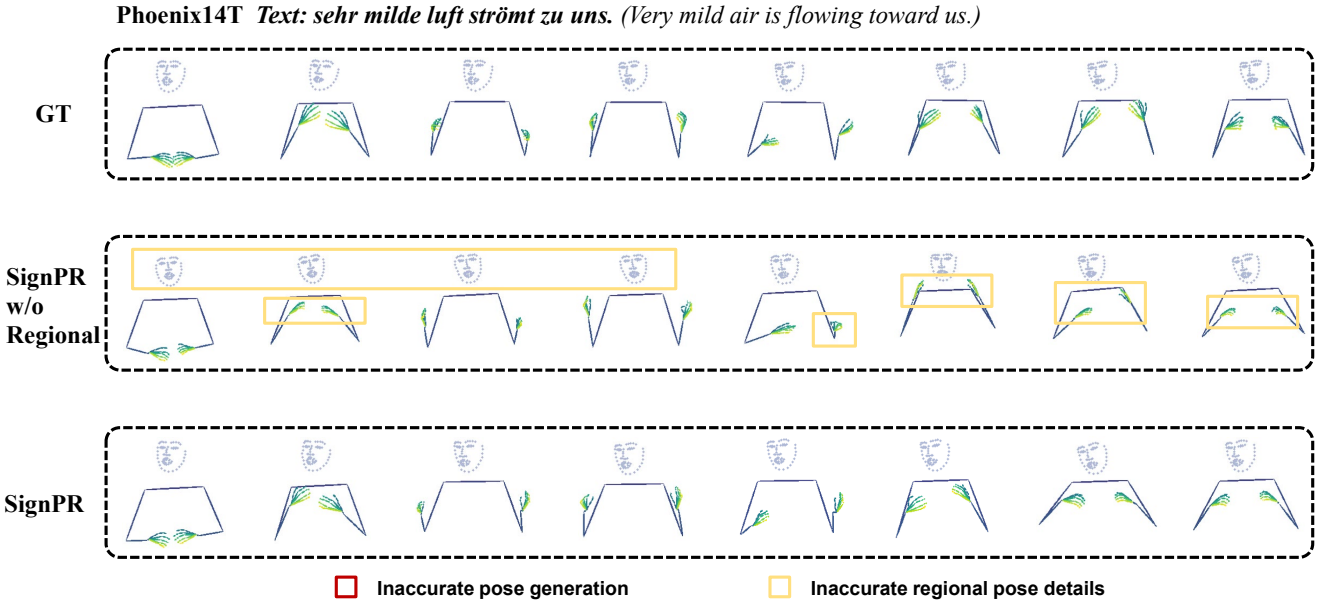


Figure 9. Qualitative results of generated poses from SignPR on Phoenix14T dataset.

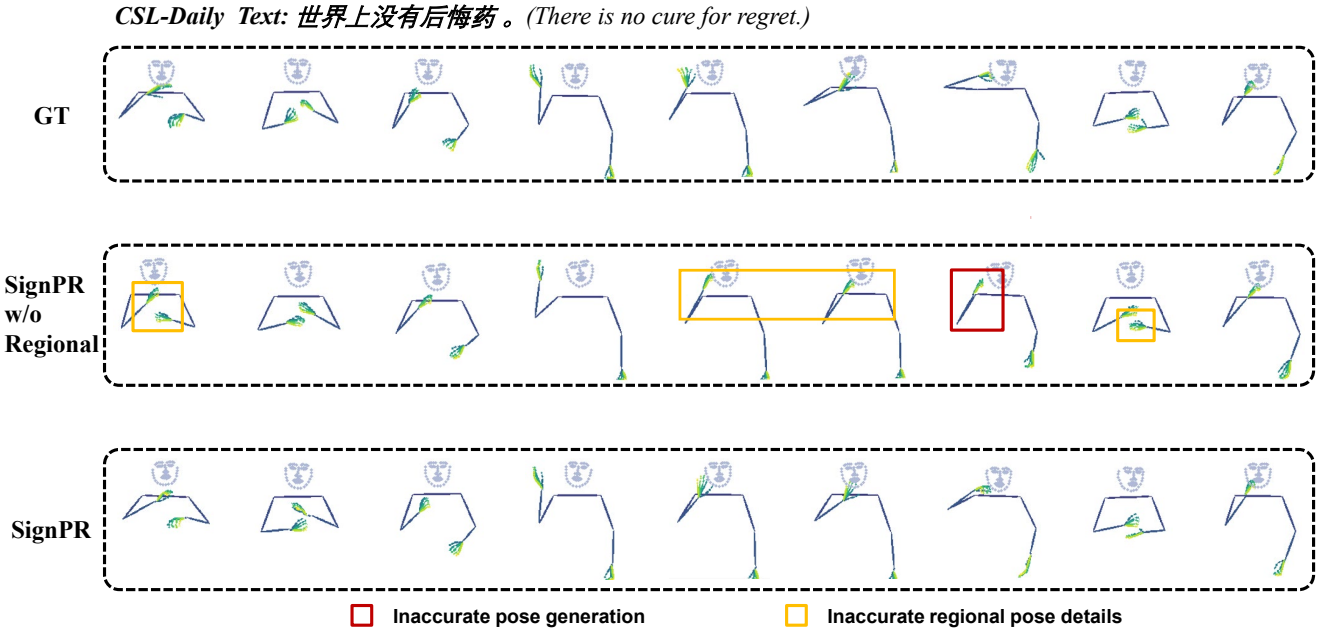


Figure 10. Qualitative results of generated poses from SignPR on the CSL-Daily dataset.

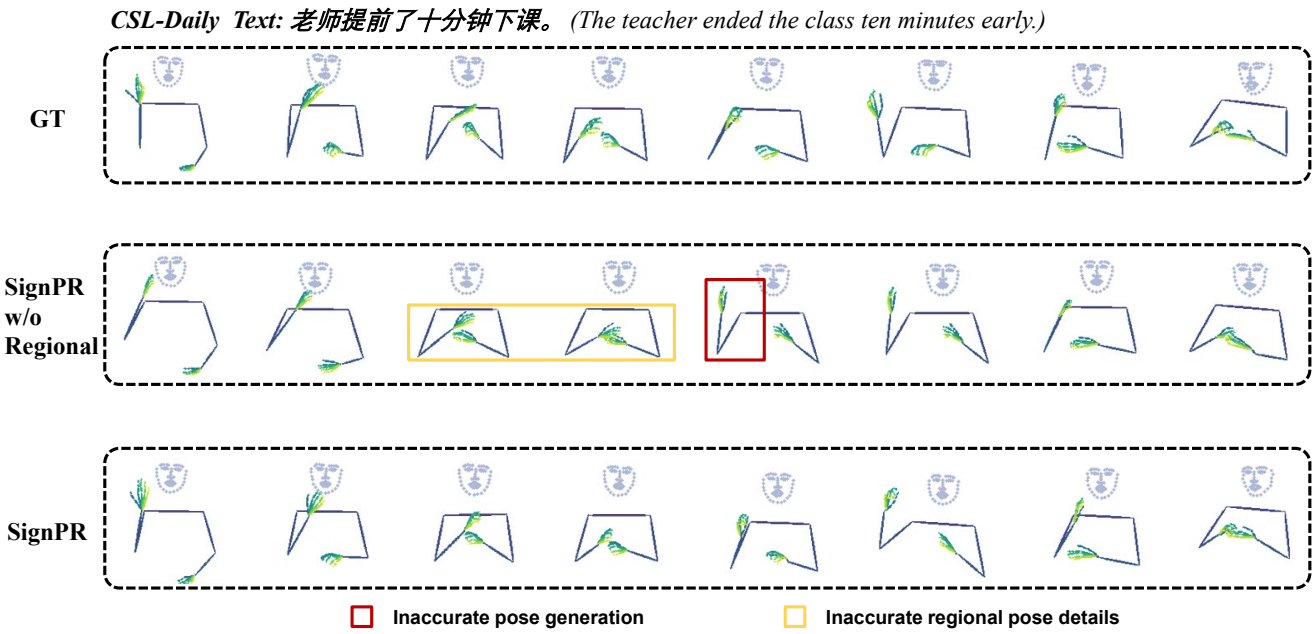


Figure 11. Qualitative results of generated poses from SignPR on the CSL-Daily dataset.

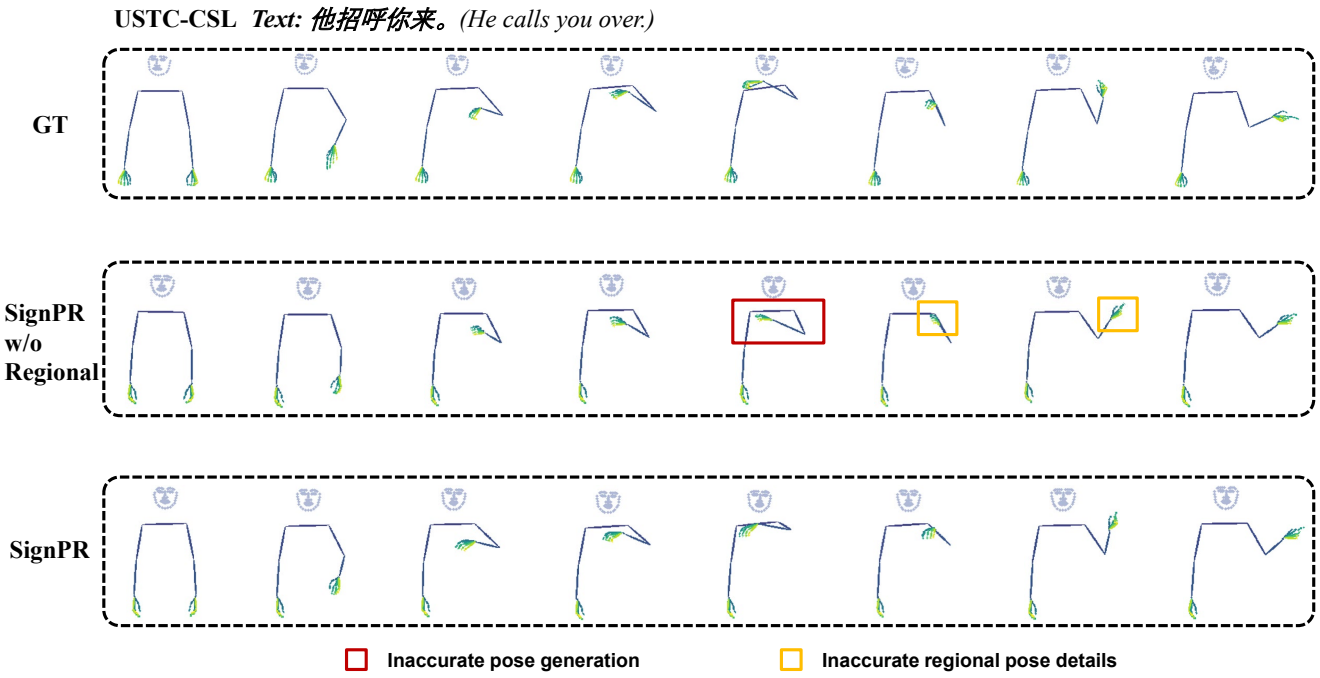


Figure 12. Qualitative results of generated poses from SignPR on USTC-CSL dataset.