Supplementary Material for "SIGNER: Temporally Grounded Sign Language Generation via Time-Resolved Conditioning"

Taeryung Lee[1], Hyeongjin Nam[2], Gyeongsik Moon[3†], and Kyoung Mu Lee[1,2†]

[1]IPAI, [2]Dept. of ECE & ASRI, Seoul National University
[3]Dept. of CSE, Korea University

1 Cross-evaluator evaluation

Previous works evaluate linguistic metrics such as WER [8], BLEU [10], and ROUGE [7] using different back-translation models, making direct comparison difficult. In particular, the reported test-set performance of the ground-truth motion varies across papers, indicating that the evaluator itself differs across methods. For example, NSA [1] and SOKE [13] rely on their own back-translation models, which are not publicly available, while G2P-DDM [12] and NAT-EA [4] use a skeleton-based evaluator. Such inconsistency makes the reported linguistic scores not directly comparable across prior works.

Our method uses a motion representation that includes not only body and hand joints but also facial expressions. Accordingly, evaluating our model with a skeleton-based back-translation model is suboptimal, as such an evaluator cannot fully assess the information preserved in face-aware motion sequences. For a fair and expressive comparison, we therefore train a unified evaluator on the same face-inclusive motion representation used in our method, and apply this evaluation setup to all compared baselines. Likewise, all comparison baselines are trained and evaluated using the same face-inclusive data representation, and generate motions with facial expressions as well (see the qualitative comparison videos in the supplementary material).

Our evaluator. We train a unified sign-to-text back-translation model on the same motion representation used throughout our framework, which includes body, hand, and facial components. Our evaluator is built by adapting the TwoStream framework [2] to our data format. It consists of two stages: a recognition model that predicts a gloss sequence from the input motion, and a translation model that reconstructs the spoken-language text.

The recognition model is composed of a pose encoder and a prediction head. For the pose encoder, we adopt an P3D-based [6] architecture. Each frame of the input pose sequence is represented using 43 joints with three modalities: 3D joint coordinates (129 dimensions), 6D joint rotations (258 dimensions), and a 13-dimensional facial expression vector. These three components are first linearly projected into separate embedding spaces, yielding 256-dimensional joint

† Corresponding authors.

Used evaluator setting	Method	CSL-Daily			Phoenix-2014T		
		WER↓	BLEU-4↑	ROUGE↑	WER↓	BLEU-4↑	ROUGE↑
Our evaluator	GT (test set)	37.86	22.44	49.20	39.32	20.73	41.51
	MotionGPT [5]	94.92	1.79	15.48	99.60	0.81	7.07
	MDM [11]	95.54	2.78	16.29	99.83	0.97	7.40
	NSA [1]	96.64	2.09	14.56	98.29	1.03	7.62
	MoMask [3]	91.51	3.79	20.12	92.01	5.15	22.80
	SOKE [13]	85.69	4.09	21.70	87.74	5.52	19.69
	G2P-DDM [12]	74.28	6.44	25.31	88.00	5.17	16.78
	NAT-EA [4]	67.80	7.87	29.17	75.65	8.79	23.04
	Ours	**55.05**	**15.60**	**39.52**	**67.16**	**11.46**	**29.39**
Previous skeleton-based evaluator (as in [4, 12])	GT (test set)	43.65	17.08	44.08	54.55	10.69	27.66
	MotionGPT [5]	94.32	1.75	15.95	95.72	2.13	10.84
	MDM [11]	92.16	2.41	16.94	91.21	3.40	15.61
	NSA [1]	91.66	2.81	18.95	94.11	3.19	14.19
	MoMask [3]	89.78	3.44	20.65	82.01	7.15	20.80
	SOKE [13]	80.94	4.17	21.89	82.86	6.81	20.33
	G2P-DDM [12]	75.01	5.96	25.37	79.66	7.87	20.65
	NAT-EA [4]	78.27	5.43	25.22	80.89	7.35	20.57
	Ours	**60.17**	**11.78**	**34.94**	**72.54**	**8.76**	**23.67**
Originally reported score (using skeleton-based evaluator as in [4, 12])	GT (test set)	Not reported			55.93	10.58	27.70
	G2P-DDM [12]	–	–	–	77.26	7.50	–
	NAT-EA [4]	–	–	–	82.01	6.66	19.43

Table S1: Comparison under different evaluator settings. We report back-translation results on CSL-Daily and Phoenix-2014T under three evaluator settings: (1) our evaluator, (2) the previous skeleton-based evaluator, and (3) originally reported scores from prior works that use the previous evaluator.

embeddings, 256-dimensional rotation embeddings, and 64-dimensional expression embeddings. They are then processed by independent temporal convolution encoders, producing modality-specific temporal features of dimensions 224, 224, and 64, respectively. Finally, these features are concatenated to obtain a 512-dimensional pose representation for each frame. This representation is further refined by a visual head composed of a linear projection layer, masked batch normalization, ReLU activation, positional encoding, dropout with rate 0.1, a position-wise feed-forward layer, and layer normalization.

For the translation stage, we initialize the translation network with pretrained mBART [9]. We use mBART-de for Phoenix-2014T and mBART-zh for CSL-Daily. To further stabilize back-translation quality, we initialize the translation network weights from a pretrained gloss-to-text checkpoint. The gloss embeddings are also initialized from pretrained mBART-based gloss embeddings. The features produced by the pose encoder are mapped into the mBART input space via a VLMapper, and the mapped features are then fed into the translation network for text generation.

We train the evaluator using Adam with $\beta = (0.9, 0.998)$ and weight decay 0.001 for 50,000 iterations with 32 mini batch size. We use a learning rate of 1×10^{-5} for the translation network and 1×10^{-4} for the remaining modules. Mixed-precision training is enabled with GradScaler.

Cross-evaluator evaluation. To further ensure transparency, we additionally evaluate all methods using the previous skeleton-based evaluator adopted in prior works such as G2P-DDM [12] and NAT-EA [4]. We report these results in the second block of Table S1. In the third block, we also list the originally re-

Method	Semantic accuracy↑	Naturalness↑
MoMask [3]	2.55	2.99
NAT-EA [4]	3.39	3.57
Spoken2Sign [14]	3.66	2.80
Ours	**4.69**	**4.77**

Table S2: Human evaluation. Participants rated generated sign motions on semantic accuracy and naturalness using a 0–5 scale. SIGNER achieves the highest score in both criteria.

ported scores of G2P-DDM and NAT-EA under their original evaluation setting. The reproduced scores remain reasonably close to the originally reported numbers, suggesting that our reproduction under the previous evaluator is reliable. More importantly, our method consistently outperforms prior approaches across evaluator choices, showing that the observed performance gains do not depend on a specific evaluator design. These results support that the advantage of our method comes from improved generation quality rather than from the choice of the back-translation model.

2 Human evaluation

We conduct a human evaluation to assess the semantic accuracy and naturalness of generated sign motions. A total of 23 participants evaluated 6 samples per method using a 0–5 scale. semantic accuracy measures how well the generated signing matches the intended meaning, while naturalness measures motion realism and transition quality. We compare SIGNER with representative baselines, including MoMask [3], NAT-EA [4], and Spoken2Sign [14]. As shown in Table S2, SIGNER achieves the highest scores in both semantic accuracy and naturalness, supporting the improvements observed in quantitative results.

References

1. Baltatzis, V., Potamias, R.A., Ververas, E., Sun, G., Deng, J., Zafeiriou, S.: Neural Sign Actors: A diffusion model for 3d sign language production from text. In: CVPR (2024)
2. Chen, Y., Zuo, R., Wei, F., Wu, Y., Liu, S., Mak, B.: Two-stream network for sign language recognition and translation. In: NeurIPS (2022)
3. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: MoMask: Generative masked modeling of 3d human motions. In: CVPR (2024)
4. Huang, W., Pan, W., Zhao, Z., Tian, Q.: Towards fast and high-quality sign language production. In: ACM MM (2021)
5. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: MotionGPT: Human motion as a foreign language. In: NeurIPS (2024)

6. Lee, T., Oh, Y., Lee, K.M.: Human part-wise 3d motion context learning for sign language recognition. In: ICCV (2023)
7. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out (2004)
8. Lin, C.Y., Och, F.J.: Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL-04). pp. 605–612 (2004)
9. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics **8**, 726–742 (2020)
10. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL (2002)
11. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: ICLR (2023)
12. Xie, P., Zhang, Q., Taiying, P., Tang, H., Du, Y., Li, Z.: G2P-DDM: Generating sign pose sequence from gloss sequence with discrete diffusion model. In: AAAI (2024)
13. Zuo, R., Potamias, R.A., Ververas, E., Deng, J., Zafeiriou, S.: Signs as Tokens: A retrieval-enhanced multilingual sign language generator. In: arXiv (2024)
14. Zuo, R., Wei, F., Chen, Z., Mak, B., Yang, J., Tong, X.: A simple baseline for spoken language to sign language translation with 3d avatars. In: ECCV (2024)