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

Abstract. Sign language generation (SLG), also known as text-to-sign generation, aims to bridge the communication gap between signers and non-signers. Unlike many other generative tasks, SLG must satisfy two fundamental linguistic constraints. First, sign language expresses meaning through a sequence of gestures aligned with word-like units called glosses, and therefore requires correct lexical ordering to preserve intended meaning. Second, each gesture should faithfully reflect the intended gloss (semantic accuracy). Despite recent progress, existing SLG methods frequently produce signs with incorrect lexical order and low semantic accuracy. A common limitation of prior approaches stems from globally fused conditioning strategies, which weaken *temporal grounding*, the temporal correspondence between glosses and their realized sign segments. This often leads to incorrect lexical order and semantically ambiguous signs. To address this limitation, we propose **SIGNER**, a **SIGN** language generation framework with tim**E-R**esolved conditioning to ensure temporal grounding, leveraging a temporal-gloss condition and local temporal fusion (LTF). SIGNER constructs a temporal-gloss condition by estimating a gloss sequence and its durations from input text, and assigning gloss semantics across the temporal dimension. We then introduce LTF, a temporally grounded fusion module that integrates the temporal-gloss condition within a constrained temporal window during denoising. By enforcing temporal locality in condition fusion, LTF preserves temporal grounding, leading to correct lexical ordering and clearer per-gloss semantics. Experiments on Phoenix-2014T and CSL-Daily demonstrate state-of-the-art performance, further supported by motion-smoothness analysis. The project page is available here.

Keywords: Sign language generation · Human motion generation

† Corresponding authors.

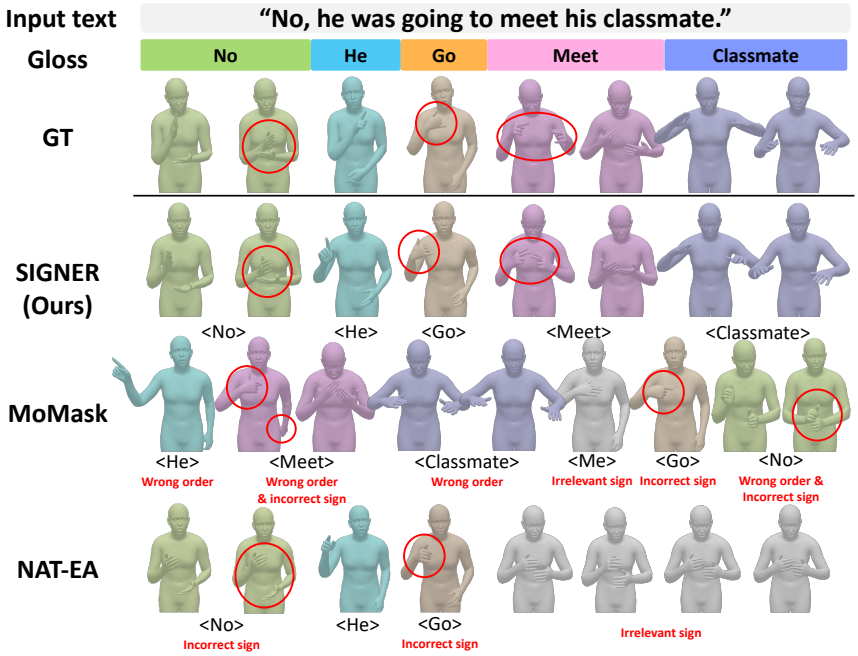


Fig. 1: Qualitative comparison. We compare the generated signs from our **SIGNER** and prior methods, MoMask [11] and NAT-EA [14], all trained on CSL-Daily. We represent the glosses with colors, and the sign segments corresponding to each gloss are painted in the same color. Human-interpreted glosses of generated signs are denoted in brackets below the signs. While the signs generated by the prior methods exhibit incorrect lexical order and include inaccurate signs, our SIGNER produces accurate signs in the correct order.

1 Introduction

Sign language is a primary communication medium for millions of deaf and hard-of-hearing (DHH) individuals worldwide. However, communication between signers and non-signers often depends on expert interpreters. Text-to-sign language generation (SLG) [3, 41, 46, 47] is a key step toward lowering this barrier, enabling applications such as automated subtitles and virtual assistants.

Unlike many other generative tasks such as text-to-image [32], SLG must satisfy strict linguistic constraints. Sign language expresses meaning through a sequence of gestures aligned with word-like linguistic units called glosses. Each gloss corresponds to a sign segment, and these segments must appear in the correct *lexical order* to preserve the intended meaning [8, 27]. In addition to lexical order, *semantic accuracy*—how faithfully each gesture reflects the intended gloss—is equally crucial for preserving clarity in sign language communication. These linguistic properties imply that SLG requires explicit *temporal grounding*,

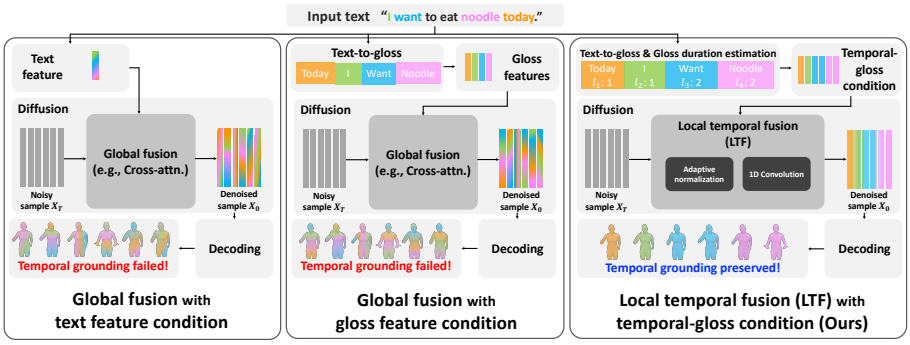


Fig. 2: Temporally grounded sign language generation. Previous SLG approaches often rely on global fusion (e.g., cross-attention), which integrates text/gloss features over the entire sequence (left/middle), without ensuring temporal grounding (i.e., temporal gloss-sign correspondence). In contrast, our method realizes time-resolved conditioning by combining temporal-gloss condition with local temporal fusion (LTF) to preserve temporal grounding (right).

namely temporal correspondence between gloss semantics and their realized sign segments (illustrated in Figure 1 by color correspondence).

Despite recent advances, existing SLG methods frequently fail to generate semantically accurate signs in correct lexical order, as illustrated in Figure 1. A common limitation among prior works is the use of global fusion [11, 32, 36, 46], where conditioning signals are shared across the entire sequence (Figure 2, left). Such global fusion provides no explicit indication of which gloss should be presented at a particular timestep, often resulting in incorrect lexical order and semantically ambiguous signs. Some works incorporate gloss sequences to better reflect lexical structure [14, 39] (Figure 2, middle). However, even in these methods, gloss-based conditions are still globally fused, which limits explicit preservation of temporal grounding and leaves the limitation of global fusion only partially addressed (Figure 1, last row).

To address these limitations, we propose **SIGNER**, a **SIGN** language generation framework with tim**E**-**R**esolved conditioning (Figure 2, right). In contrast to prior methods based on global fusion, which share conditioning signals across the entire sequence, SIGNER preserves the temporal correspondence between glosses and their realized sign segments (*i.e.,* temporal grounding). By explicitly maintaining temporal grounding during generation, SIGNER prevents gloss information from being mixed across time, which often causes incorrect lexical order and semantically ambiguous signs.

SIGNER implements time-resolved conditioning by combining a temporal-gloss condition with local temporal fusion (LTF). We construct the temporal-gloss condition by deriving a gloss sequence from the input text, estimating gloss durations, and repeating each gloss embedding for its estimated duration to form a time-resolved sequence. Our key contribution is LTF, which integrates this

condition into diffusion denoising within a local temporal window. In contrast to global fusion that integrates the condition globally across time and weakens time-specific guidance, LTF preserves temporal grounding by applying the condition locally, which enforces the correct lexical order. Moreover, this locality constraint prevents semantic overmixing across adjacent segments, improving semantic accuracy.

We evaluate SIGNER on Phoenix-2014T [4] and CSL-Daily [45], where it substantially outperforms prior SLG methods on back-translation-based metrics. Beyond automatic evaluation, we additionally conduct a motion-smoothness analysis (e.g., peak jerk statistics) to validate overall motion quality. Our contributions are summarized as follows.

- We propose **SIGNER**, a sign language generation framework via time-resolved conditioning to enforce lexical order of generated signs.
- We introduce local temporal fusion (LTF), which integrates the temporal-gloss condition locally during denoising to ensure the temporal grounding.
- LTF also enhances semantic accuracy by mitigating overmixing of gloss semantics across time.

2 Related works

Sign language generation (SLG). SLG synthesizes sign gesture sequences from spoken-language text, enabling machine-mediated communication for deaf and hard-of-hearing individuals [3, 16, 33, 34, 41, 42, 46, 47]. Recent methods span diverse generation paradigms, including diffusion-based generation [3], autoregressive token generation [41, 46], and retrieval-and-concatenation pipelines [47]. Despite progress, two limitations remain central: generating signs in correct lexical order and maintaining semantic accuracy at the segment level. A common cause is global fusion, where text or gloss features are shared across the entire sequence, weakening temporal grounding and leading to cross-time confusion of lexical cues and temporally mixed semantics. Several works incorporate gloss features [39] or temporal-gloss conditions [14] to better reflect lexical structure. However, such conditions are still globally fused into generation process via cross-attention, leaving these limitations not fully resolved. Retrieval-based pipelines such as Spoken2Sign [47] can enforce lexical order by sequentially connecting retrieved sign segments, but they often yield unnatural transitions due to limited modeling of temporal dependencies across segment boundaries. In contrast, our approach explicitly enforces temporal grounding through time-resolved conditioning, combining a temporal-gloss condition with LTF. This directly addresses the limitations of previous approaches using global fusion, improving lexical order and semantic accuracy.

Human motion generation. Human motion generation aims to synthesize temporally coherent body movements from text descriptions [1,6,11,12,17–20,30, 31, 35, 36, 38, 43, 44]. ACTOR [30] learns action-conditioned motion embeddings with a Transformer VAE. MDM [36] applies diffusion in motion space for text-to-motion generation. MotionGPT [17] models motion as discrete tokens for

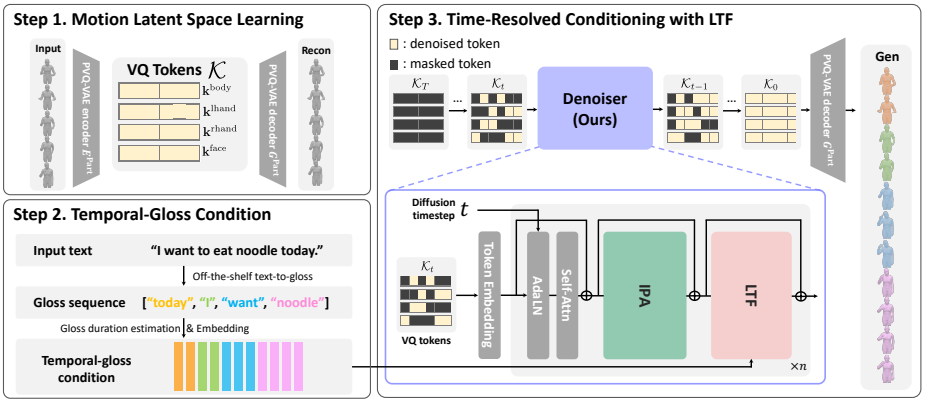


Fig. 3: Overview. Our method aims to model temporal correspondence between the glosses and their associated sign segments (temporal grounding). We first learn a motion latent space using PVQ-VAE (top-left). Next, we construct temporal-gloss condition from the input text (bottom-left), providing time-resolved conditioning signals. Finally, SIGNER generates sign language sequences by integrating temporal-gloss conditions within localized temporal windows via local temporal fusion (LTF), enabling accurate temporal grounding.

unified pretraining. MoMask [11] uses masked modeling over a hierarchical VQ representation. While these methods can be adapted to SLG, sign language imposes additional linguistic constraints such as lexical order and gloss-level semantic specificity. Our temporally-localized conditioning is designed to better respect these constraints during generation.

3 Preliminary

Motion Representation. We utilize the human motion representation defined on 43 joints of upper-body and hands, consisting of 3D joint position, 6D rotation and 3D velocity. We also leverage facial expression parameters and jaw poses. Therefore, our pose representation consists of $43 \times (3+6+3) + 10 + 3 = 529$ dimensions.

VQ-VAE. VQ-VAE [37, 43] comprises an encoder E, a decoder G, and a codebook $\mathcal{Z} = \{\mathbf{z}_k \in \mathbb{R}^d\}_{k=1}^K$, where K is the number of codebook entries and d is the dimensionality of each code vector. Given an input motion $\mathbf{X} \in \mathbb{R}^{L \times D}$, where L and D denote the sequence length and feature dimension, respectively, the encoder maps it to a latent representation $\mathbf{Z}^e = E(\mathbf{X}) \in \mathbb{R}^{L' \times d}$. The length of the latent sequence is defined as $L' = \lfloor L/r \rfloor$, where r denotes the temporal reduction rate of the encoder E. We denote the encoder outputs and quantized latent representation as $\mathbf{Z}^e$ and $\mathbf{Z}^q$, respectively. Each latent vector $\mathbf{Z}_i^e$ at temporal step i is quantized to the nearest code $\mathbf{z}_k$:

$$\mathbf{Z}_i^q = \mathbf{z}_k, \quad \text{where } k = \arg\min_{k'} \|\mathbf{Z}_i^e - \mathbf{z}_{k'}\|_2. \tag{1}$$

The decoder reconstructs the motion from the quantized latent codes $\hat{\mathbf{X}} = G(\mathbf{Z}^q)$. The VQ-VAE is trained to minimize the following loss function:

$$\mathcal{L}_{\text{VQ}} = \|\mathbf{X} - \hat{\mathbf{X}}\|_2^2 + \|\text{sg}(\mathbf{Z}^e) - \mathbf{Z}^q\|_2^2 + \beta\|\mathbf{Z}^e - \text{sg}(\mathbf{Z}^q)\|_2^2, \tag{2}$$

where $\text{sg}(\cdot)$ is the stop-gradient operation, and β is a loss weight.

Discrete diffusion. Discrete diffusion [10] operates over discrete tokens, a finite set of codebook indices, obtained from a pre-trained VQ-VAE. Unlike continuous diffusion models [13], which add Gaussian noise to real-valued inputs, this approach defines a Markov diffusion process over discrete tokens.

Let $\mathbf{k}_0 \in \{1, \ldots, K\}^{L'}$ denote the original sequence of discrete tokens from the codebook. The forward diffusion process gradually corrupts $\mathbf{k}_0$ over T timesteps by randomly replacing tokens, defined by a transition matrix Q_t:

$$q(\mathbf{k}_t \mid \mathbf{k}_{t-1}) = \mathbf{v}^\top(\mathbf{k}_t) Q_t \mathbf{v}(\mathbf{k}_{t-1}), \tag{3}$$

where $\mathbf{v}(\cdot)$ denotes the one-hot representation of a discrete token, and $Q_t \in \mathbb{R}^{K \times K}$ is a stochastic matrix with rows summing to 1. The transition matrix is constructed using hyperparameters α_t, β_t, and γ_t, which control the probabilities of token retention, replacement with random tokens, and masking, respectively. At the final timestep T, all tokens are fully masked, and the masked sequence serves as the starting point for the reverse denoising process, analogous to sampling from a standard Gaussian distribution in continuous DDPM frameworks.

The goal of the reverse process is to recover $\mathbf{k}_0$ from a noisy $\mathbf{k}_t$. This is done by estimating the reverse transition:

$$p_\theta(\mathbf{k}_{t-1} \mid \mathbf{k}_t, \mathbf{y}) = \sum_{\tilde{\mathbf{k}}_0} q(\mathbf{k}_{t-1} \mid \mathbf{k}_t, \tilde{\mathbf{k}}_0)\, p_\theta(\tilde{\mathbf{k}}_0 \mid \mathbf{k}_t, \mathbf{y}), \tag{4}$$

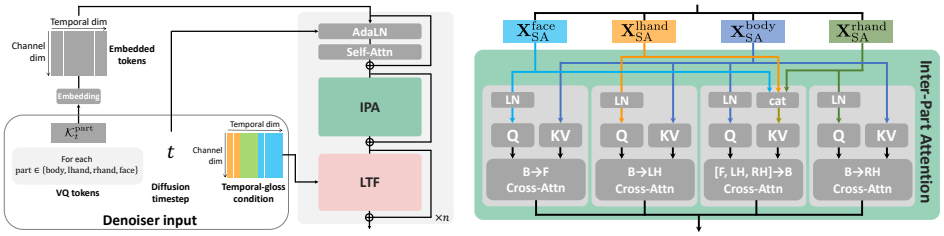
where $\tilde{\mathbf{k}}_0$ is the predicted denoised discrete token sequence, and $\mathbf{y}$ is a conditional input (e.g., text prompt). The denoiser model, which can be designed for the specific purpose such as sign language generation, predicts the initial state $p_\theta(\tilde{\mathbf{k}}_0 \mid \mathbf{k}_t, \mathbf{y})$. Denoiser is trained to estimate the posterior transition distribution $q(\mathbf{k}_{t-1} \mid \mathbf{k}_t, \mathbf{k}_0)$ with the variational lower bound [10].

4 SIGNER

In this section, we describe SIGNER, which implements time-resolved conditioning by using temporal-gloss condition and local temporal fusion (LTF) (Figure 3). We first introduce part-aware motion latents learned by PVQ-VAE, then construct the temporal-gloss condition, and finally describe time-resolved conditioning with LTF.

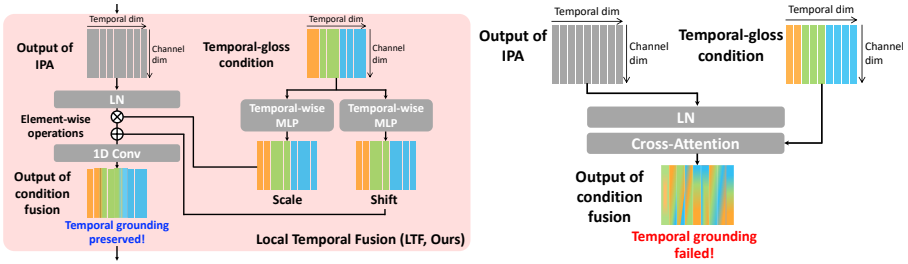
4.1 Motion latent space learning

As illustrated in Figure 3, we learn the motion latent space using PVQ-VAE to enhance the expressiveness of hand and face, inspired by prior works [5,25,39,46].



(a) Denoiser. **(b)** Inter-Part Attention (IPA).

Fig. 4: Model architecture. We illustrate the detailed architecture of our proposed denoiser and inter-part attention (IPA) module. (a) In our model, each body part—body, left hand, right hand, and face—is encoded independently without inter-part interaction, except within the IPA module. (b) Cross-attention between different body parts in the IPA block facilitates more coordinated movements of articulations, leading to a better quality of generated sign language.



(a) Local Temporal Fusion (LTF, **Ours**). **(b)** Cross-attention.

Fig. 5: Comparison of condition fusion methods. (a) Our LTF consists of AdaLN and 1D convolution to preserve the temporal grounding within local temporal context. (b) Cross-attention, which fails to maintain the temporal grounding.

Following the notations in previous section, each input motion sequence $\mathbf{X}$ is decomposed into $\mathbf{X}^{\text{body}}$, $\mathbf{X}^{\text{lhand}}$, $\mathbf{X}^{\text{rhand}}$, and $\mathbf{X}^{\text{face}}$. We denote the input motion sequence of each part by $\mathbf{X}^{\text{part}}$, where part $\in \{\text{body}, \text{lhand}, \text{rhand}, \text{face}\}$.

The input motion sequence of each part $\mathbf{X}^{\text{part}}$ is processed by a dedicated 1D convolutional encoder E^{part} and decoder G^{part}, with a corresponding codebook of each part $\mathcal{Z}^{\text{part}} = \{\mathbf{z}_k^{\text{part}} \in \mathbb{R}^{\text{d}}\}_{k=1}^{K}$. Each input $\mathbf{X}^{\text{part}}$ is encoded as $\mathbf{Z}^{e,\text{part}} = E^{\text{part}}(\mathbf{X}^{\text{part}})$, then quantized using the codebook into $\mathbf{Z}^{q,\text{part}}$. The quantized feature is decoded as $\hat{\mathbf{X}}^{\text{part}} = G^{\text{part}}(\mathbf{Z}^{q,\text{part}})$. Here, $\mathbf{Z}^{e,\text{part}}$ and $\mathbf{Z}^{q,\text{part}}$ denote the encoder output and the quantized latent of each part, respectively. The training objective minimizes the loss across all parts, $\mathcal{L} = \mathcal{L}^{\text{body}} + \mathcal{L}^{\text{lhand}} + \mathcal{L}^{\text{rhand}} + \mathcal{L}^{\text{face}}$, with each term defined in Equation 2.

4.2 Temporal-gloss condition

As illustrated in Figure 3 (bottom left), we first convert the input text into a gloss sequence $g = \{g_j\}_j$ using an off-the-shelf text-to-gloss model used in [47]. For each gloss g_j, we extract an embedding $c_j \in \mathbb{R}^{D_{\text{cond}}}$ using a pre-trained

mBART [23]. A learned duration estimator (MLP) predicts token-level durations l_j from c_j, and we define the generation length as $L_g = \sum_j l_j$. We then construct an intermediate sequence $\bar{S} \in \mathbb{R}^{L_g \times D_{\text{cond}}}$ by repeating each c_j for l_j steps and concatenating them in order. Finally, we project $\bar{S}$ to the denoiser feature dimension to obtain $S = \Phi(\bar{S}) \in \mathbb{R}^{L_g \times D_{\text{feat}}}$. During training and inference, we use token sequences of length L_g for both the motion tokens and the temporal-gloss condition.

4.3 Time-resolved conditioning with LTF

In this section, we describe the temporally grounded sign language generation process with local temporal fusion (LTF). First, we describe the overall architecture. Then we introduce inter-part attention (IPA) to enable coordinated movements across articulators, which is critical when separately encoding each body part with PVQ-VAE. We also present LTF to promote temporal localization during condition fusion.

Architecture. Figure 4a visualizes the architecture of the proposed denoiser. Each body part—body, left hand, right hand, and face—is forwarded independently without inter-part interaction, except within the IPA module. To be specific, for each part, we first represent discrete tokens $\mathbf{k}^{\text{part}} \in \{1, \dots, K\}^{L_g}$ as a vector of codebook indices. The discrete tokens are formed into embedded tokens $\mathbf{X}^{\text{part}}_{\text{embed}} \in \mathbb{R}^{L_g \times D_{\text{feat}}}$. D_{feat} denotes the channel dimension of the feature, and is maintained throughout the denoiser forward pass. Then, we integrate the diffusion timestep t into the embedded tokens $\mathbf{X}^{\text{part}}_{\text{embed}}$ using adaptive layer normalization (AdaLN) [15], resulting in $\mathbf{X}^{\text{part}}_{\text{time}}$. This is passed through a self-attention (SA) layer followed by a residual connection, yielding $\mathbf{X}^{\text{part}}_{\text{SA}} = \text{SA}(\mathbf{X}^{\text{part}}_{\text{time}}) + \mathbf{X}^{\text{part}}_{\text{embed}}$. Next, $\mathbf{X}^{\text{part}}_{\text{SA}}$ is processed by the inter-part attention (IPA) module with a residual connection, producing $\mathbf{X}^{\text{part}}_{\text{IPA}} = \text{IPA}(\mathbf{X}^{\text{part}}_{\text{SA}}) + \mathbf{X}^{\text{part}}_{\text{SA}}$. Finally, the features are fused with the temporal-gloss condition S in LTF, resulting in the output $\mathbf{X}^{\text{part}}_{\text{LTF}}$.

Inter-part attention (IPA). Figure 4b illustrates the architecture of our proposed IPA module. It is designed to enhance the denoising process by enabling coordinated movements across articulators through inter-part feature integration. Each of body features (body, hands, face) is first normalized with LayerNorm [2] and then used as a query in cross-attention with other body parts. Specifically, the body features attend to a concatenation of the face and hand features, while each hand and face attends to the body features. The model captures dependencies between body parts through this cross-attention mechanism, producing refined features for the body and both hands.

Local temporal fusion (LTF). Figure 5a shows LTF for temporal grounding, which integrates the temporal-gloss condition within a local temporal window. LTF anchors the gloss semantics to their corresponding token index from the temporal-gloss condition, and aggregates nearby context to produce smooth and natural transitions.

Given a temporal-gloss condition $S \in \mathbb{R}^{L_g \times D_{\text{feat}}}$, two MLPs produce per-token scale and shift, $u, v \in \mathbb{R}^{L_g \times D_{\text{feat}}}$. We modulate the denoiser features with

adaptive normalization and then apply a temporal 1D convolution:

$$\mathbf{X}_{\text{LTF}}^{\text{part}} = \text{Conv1D}\big(\text{LN}(\mathbf{X}_{\text{IPA}}^{\text{part}}) \odot (1+u) + v\big)\,,$$

where the convolution aggregates local temporal context within a fixed window. Through this design, AdaLN ensures that each gloss condition is aligned to its corresponding token index, while the 1D convolution allows the model to consider local context for coherent generation. In contrast to global fusion (e.g., cross-attention over all gloss tokens), LTF confines conditioning to local neighborhoods, improving temporal grounding and transition coherence.

Training objective. We follow the standard discrete diffusion objective in VQ-diffusion [10]. Given a clean token sequence $\mathbf{k}_0$ and its noised version $\mathbf{k}_t$ at diffusion step t, the denoiser predicts the posterior over the original tokens conditioned on the temporal-gloss condition $p_\theta(\mathbf{k}_0 \mid \mathbf{k}_t, t, S)$. We train the model by minimizing the negative log-likelihood of the ground-truth tokens:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\,\mathbf{k}_0,\,\mathbf{k}_t}\Big[-\log p_\theta(\mathbf{k}_0 \mid \mathbf{k}_t, t, S)\Big],$$

where t is sampled uniformly and $\mathbf{k}_t$ is obtained by the forward noising process.

5 Experiment

5.1 Protocols

Datasets. We use two large-scale sign language datasets: CSL-Daily [45] and PHOENIX-2014T [4], which contain approximately 20K and 8K samples, respectively, in Chinese and German sign languages. Our models are trained with a maximum motion length of 180 frames [36,43], and all evaluations are conducted on the test set.

Implementation details. We implement our framework using PyTorch [29]. Input text is converted into gloss sequences via an off-the-shelf model used in [47], and encoded into temporal-gloss condition with mBART [23] embeddings. We adopt the same hyperparameters for PVQ-VAE as used in T2M-GPT [43]. Our discrete diffusion model follows the configuration of VQ-Diffusion [10], except that we use 50 timesteps. The denoiser consists of 14 layers with an inner dimension of 512 and 16 attention heads. LTF is implemented with 1D convolutional network of stride 1 and kernel size 3. Gloss duration estimator consists of 2 MLP layers with 512 dimensions. We train the models using AdamW [24] with a learning rate of 1×10^{-4} and weight decay of 4.5×10^{-2}. PVQ-VAE and the denoiser are both trained for 30K iterations, with a mini-batch size of 56 on a single NVIDIA RTX 4090 GPU.

Evaluation metrics. Following prior work [3, 46, 47], we evaluate generated signs using metrics for motion accuracy and linguistic fidelity. For motion-level evaluation, we use dynamic time warping with joint position error (DTW-JPE) [26],

Condition fusion	DTW-JPE↓			Back translation		
	Body	Hand	All	WER↓	BLEU-4↑	ROUGE↑
Cross Attention	0.251	1.258	0.832	84.90	5.81	24.87
AdaLN + FC	0.249	1.217	0.807	72.52	7.85	27.90
AdaLN + 1D Conv (LTF, Ours)	**0.242**	**1.189**	**0.788**	**55.05**	**15.60**	**39.52**

Table 1: Effectiveness of LTF. We validate the contribution of LTF by comparing different condition fusion methods. Our method appears in the last row.

Attention			DTW-JPE↓			Back translation		
B→H	H→B	B→F	Body	Hand	All	WER↓	BLEU-4↑	ROUGE↑
✗	✗	✗	0.248	1.197	0.796	61.31	13.27	36.42
✓	✗	✗	0.247	1.195	0.794	57.01	13.76	37.74
✗	✓	✗	**0.240**	1.190	**0.788**	58.74	14.52	38.13
✗	✗	✓	0.242	1.191	0.790	58.96	13.66	37.38
✓	✗	✓	0.243	1.194	0.792	55.31	15.06	38.95
✗	✓	✓	0.248	1.206	0.801	57.77	13.41	37.62
✓	✓	✗	0.247	**1.189**	0.790	57.34	13.59	37.67
✓	✓	✓	0.242	**1.189**	**0.788**	**55.05**	**15.60**	**39.52**

Table 2: Effectiveness of IPA. We validate the contribution of IPA by testing various combinations of attention flows: Body→Hands (B→H), Hands→Body (H→B), and Body→Face (B→F). Our method appears in the last row.

which measures distances between temporally aligned joint trajectories of generated and reference motions. For linguistic fidelity, we adopt a back-translation pipeline [7]: we report word error rate (WER) [22] for sign-to-gloss, and BLEU-4 [28] and ROUGE [21] for sign-to-text.

Since our method uses a new motion representation including face, evaluating our method with prior evaluators that use skeleton-only inputs can be confounded (limiting evaluation of facial expression). Therefore, for fair comparison, we re-evaluate all methods under a unified setup: we train all models to use the same motion representation (including face as part of the input/output feature), and evaluate them with a single back-translation model trained on the same representation. Details of the unified evaluator are provided in the supplementary material; for transparency, we additionally report cross-evaluator results using the prior evaluator in the supplementary material.

5.2 Ablation Study

Local temporal fusion (LTF). Table 1 shows that time-resolved conditioning with LTF yields substantial performance gains. Global fusion via cross-attention (row 1) yields limited generation performance, as it allows each motion timestep to aggregate information from all gloss tokens, which can blur temporal grounding. In contrast, our LTF (row 3) restricts conditioning to a local temporal

Text-to-gloss method	DTW-JPE↓			Back translation		
	Body	Hand	All	WER↓	Bleu-4↑	ROUGE↑
GT gloss	0.234	1.145	0.768	54.70	15.69	40.33
mT5 finetuned	**0.237**	**1.151**	**0.764**	58.70	14.87	38.92
M2M finetuned	0.250	1.215	0.806	67.63	14.79	39.08
mBART finetuned	0.252	1.204	0.801	70.48	14.78	38.44
Ours (w. off-the-shelf in [47])	0.242	1.189	0.788	**55.05**	**15.60**	**39.52**

(a) Robustness to text-to-gloss models.

Duration variation	DTW-JPE↓			Back translation		
	Body	Hand	All	WER↓	Bleu-4↑	ROUGE↑
GT (test set) duration	0.201	0.448	0.345	57.76	14.44	38.94
Estimated duration + 4	0.247	1.204	0.799	56.60	15.18	38.71
Estimated duration - 4	0.251	1.211	0.804	59.25	14.47	36.97
Estimated duration + $t \sim U(-4, 4)$	0.254	1.219	0.810	60.74	14.06	35.53
Estimated duration (Ours)	**0.242**	**1.189**	**0.788**	**55.05**	**15.60**	**39.52**

(b) Robustness to gloss duration variation.

Table 3: Robustness of our system. We show the robustness of our method on the choice of text-to-gloss models and the variation of gloss duration. Our method appears in the last row.

Setting	Method	CSL-Daily						Phoenix-2014T					
		DTW-JPE↓			Back translation			DTW-JPE↓			Back translation		
		Body	Hand	All	WER↓	BLEU-4↑	ROUGE↑	Body	Hand	All	WER↓	BLEU-4↑	ROUGE↑
Gloss free	MotionGPT [17]	0.939	2.258	1.700	94.92	1.79	15.48	1.709	3.424	2.699	99.60	0.81	7.07
	MDM [36]	0.351	1.754	1.161	95.54	2.78	16.29	0.855	2.819	1.988	99.83	0.97	7.40
	NSA [3]	0.347	1.691	1.123	96.64	2.09	14.56	0.826	2.406	1.738	98.29	1.03	7.62
	MoMask [11]	0.376	1.565	1.062	91.51	3.79	20.12	0.864	2.296	1.690	92.01	5.15	22.80
	SOKE [46]	0.256	1.283	0.849	85.69	4.09	21.70	0.231	1.102	0.733	87.74	5.52	19.69
Gloss based	G2P-DDM [39]	0.249	1.248	0.825	74.28	6.44	25.31	0.236	1.138	0.757	88.00	5.17	16.78
	NAT-EA [14]	0.532	1.598	1.147	67.80	7.87	29.17	0.453	1.475	1.043	75.65	8.79	23.04
	Ours	**0.242**	**1.189**	**0.788**	**55.05**	**15.60**	**39.52**	**0.229**	**1.096**	**0.729**	**67.16**	**11.46**	**29.39**

Table 4: Comparison to state-of-the-art methods. We report motion errors and back-translation results under a unified evaluation setting on the CSL-Daily and Phoenix-2014T datasets. Methods are grouped into gloss-free and gloss-based settings for clearer comparison. Our SIGNER outperforms previous methods on both datasets.

window, and thereby preserves temporal grounding. This localized fusion leads to accurate signs in the correct lexical order. Replacing the temporal convolution with a fully connected layer (row 2) consistently degrades performance, confirming that local temporal aggregation is critical for incorporating nearby context and producing smooth transitions. Overall, these results indicate that time-resolved conditioning in diffusion process effectively addresses the limitations of global fusion used in prior methods.

Inter-part attention (IPA). Table 2 shows that IPA improves sign language generation performance notably. As illustrated in Figure 4b, IPA integrates inter-part dependencies through cross-attention flows between body and hands (B→H and H→B) as well as body-to-face attention (B→F). Individually enabling B→H or H→B attention (second and third rows) yields consistent gains over the baseline without IPA (first row), highlighting the importance of coordinating body and hand articulations. Incorporating B→F attention further improves performance (fourth and fifth rows), indicating that facial cues provide complementary information beyond body–hand interactions. Notably, enabling all attention flows (B→H, H→B, and B→F; last row) achieves the best overall results. Based on these findings, we adopt this full IPA configuration as our final design.

Robustness on text-to-gloss and duration sampling. Table 3 evaluates the robustness of our framework to variations in both the text-to-gloss model and the predicted sign duration. The left block reports results using alternative text-to-gloss models (fine-tuned mT5 [40], M2M [9], and mBART [23]). Performance

Model	Peak velocity (↓)	Peak jerk (↓)
NSA [3]	0.75±0.20	1.75±0.59
MoMask [11]	0.44±0.13	0.63±0.31
Spoken2Sign [47]	0.88±1.11	1.70±2.14
NAT-EA [14]	0.43±0.10	0.69±0.31
Ours	**0.38±0.10**	**0.35±0.12**

Table 5: Peak jerk and peak velocity. We report peak velocity and peak jerk which indicates the naturalness and smoothness of generated motion. Results are reported in normalized units.

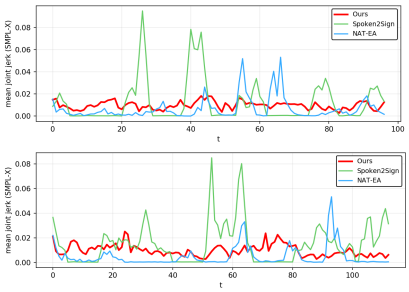


Fig. 6: Jerk curves. We visualize mean joint jerk over time for two samples. Lower peak height indicates smoother transitions.

remains strong across different text-to-gloss choices, indicating that SIGNER is not overly sensitive to the specific text-to-gloss model.

The right block (Table 3 (b)) tests robustness to perturbations around the estimated sign duration. We test robustness by perturbing the estimated duration for each gloss by +4, -4, and a random offset sampled from $U(-4, 4)$. Across these settings, performance remains stable, suggesting that SIGNER is not overly sensitive to moderate errors or noise in the predicted duration. Given that the average gloss duration in CSL-Daily is 16 frames, these perturbations correspond to approximately ±25% variation, supporting the temporal robustness of our method.

5.3 Comparison to state-of-the-art methods

Table 4 demonstrates that our proposed SIGNER substantially outperforms previous methods on both Phoenix-2014T [4] and CSL-Daily [45].

A key takeaway is that how conditions are fused across time is critical. On CSL-Daily (left block), SIGNER achieves state-of-the-art performance among all prior methods. Compared to global-fusion approaches such as MoMask [11] and SOKE [46], SIGNER outperforms by a large margin. Methods that incorporate gloss-level cues, including G2P-DDM [39] and NAT-EA [14], generally outperform text-only baselines, yet SIGNER still yields substantially higher scores. Notably, the performance gap between our SIGNER and NAT-EA [14], which also leverages a temporal-gloss condition, suggests that using temporal-gloss information alone is insufficient to guarantee correct lexical ordering and semantic accuracy. Rather, the conditioning mechanism must preserve temporal grounding during generation. Similarly, on Phoenix-2014T (right block), SIGNER delivers consistent improvements over existing methods across all evaluation metrics. These gains across datasets further support that time-resolved conditioning by LTF enhances sign generation quality by ensuring temporal grounding.

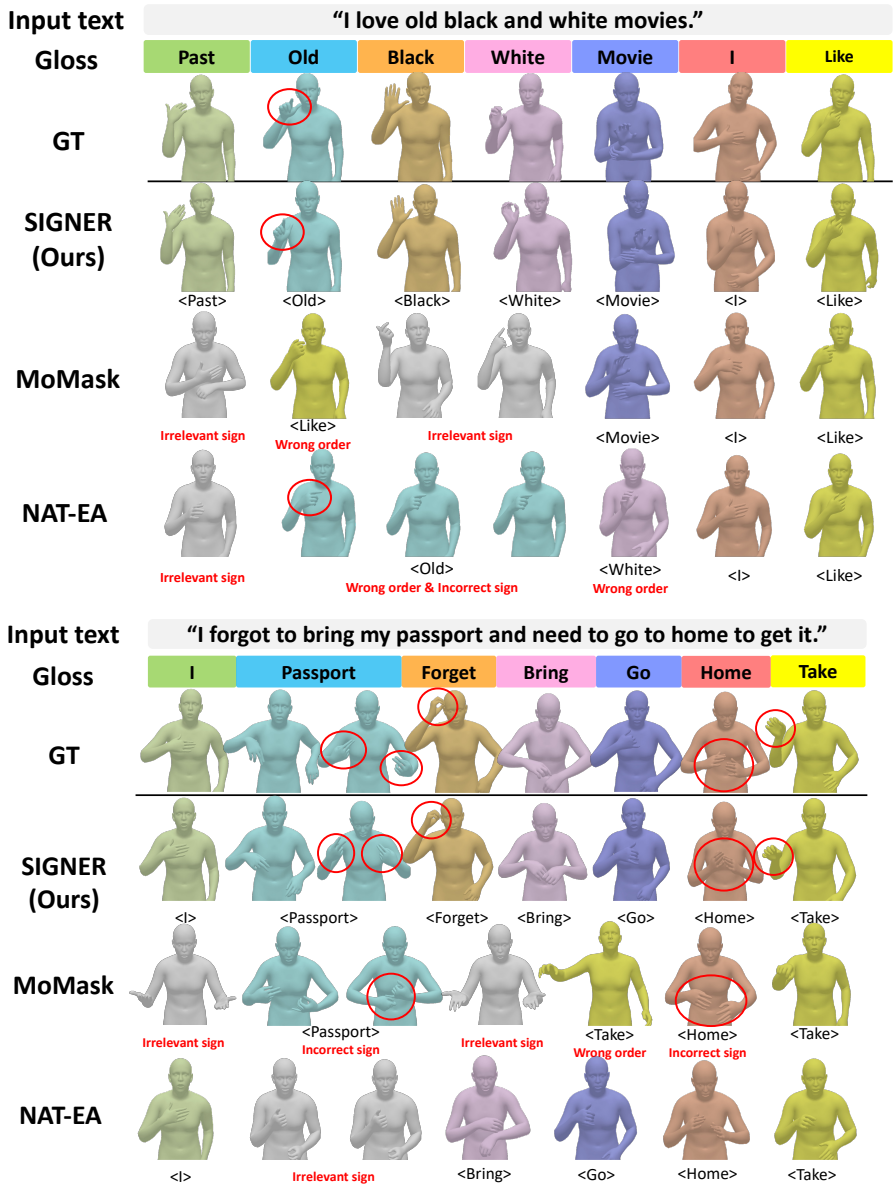


Fig. 7: Qualitative result. We present the qualitative comparison of our SIGNER with GT, MoMask [11], and NAT-EA [14], all trained on CSL-Daily.

Motion smoothness and transition quality. We evaluate motion smoothness and transition quality using peak velocity and peak jerk, where lower values indicate smoother motions. As shown in Table 5, SIGNER achieves the lowest

peak velocity and peak jerk among the compared methods. Figure 6 provides qualitative evidence: our jerk curves exhibit lower peaks, whereas baselines show pronounced peaks around transitions. We include Spoken2Sign [47] as a representative retrieval-and-concatenation baseline to contextualize transition quality.
Qualitative result. Figure 7 shows that our SIGNER generates accurate sign language in correct lexical order compared with MoMask [11] and NAT-EA [14]. Each gloss is represented by its own colors, and the corresponding sign segment is shown in the same color. Human-interpreted glosses are denoted inside brackets. Compared with our method, MoMask conditions the denoiser on a text feature via global fusion, which does not preserve temporal grounding; as a result, it often produces signs with incorrect lexical order and semantically mismatched segments. NAT-EA leverages a temporal-gloss condition, but it still fuses the condition globally, so temporal grounding is only partially preserved. As a result, NAT-EA still generates signs with incorrect lexical order and semantically mismatched segments.

6 Conclusion

Summary. We presented SIGNER, a temporally grounded sign language generation framework that addresses two central limitations of prior SLG methods: incorrect lexical ordering and low semantic accuracy. We identified globally fused conditioning as a key cause, as it weakens temporal grounding—the temporal correspondence between glosses and their realized sign segments—by mixing linguistic cues across time. To preserve temporal grounding during generation, SIGNER implements time-resolved conditioning with a temporal-gloss condition and local temporal fusion (LTF), which injects gloss semantics within localized temporal windows during diffusion denoising. Experiments on Phoenix-2014T and CSL-Daily demonstrate that SIGNER achieves state-of-the-art performance under a unified evaluation setup, showing that our proposed time-resolved conditioning efficiently addresses the challenge of lexical ordering and semantic accuracy.
Limitation. While SIGNER demonstrates strong performance in sign language generation, it relies on off-the-shelf text-to-gloss models. Despite this dependency, Table 3(a) shows that SIGNER remains robust across different text-to-gloss models. However, to train SIGNER with a new sign language dataset on a different language system, a text-to-gloss model must still be trained to extract the necessary gloss annotations. Future work could explore jointly optimizing text-to-gloss and sign generation within a unified framework to alleviate this dependency.

Acknowledgements

This work was supported in part by the IITPgrants [No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University), No. RS-2025-02303870, No.2022-0-00156] funded by the Korea government (MSIT).

References

1. Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: TEACH: Temporal Action Compositions for 3D Humans. In: 3DV (2022)
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. In: arXiv (2016)
3. 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)
4. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: CVPR (2018)
5. Chen, C., Zhang, J., Lakshmikanth, S.K., Fang, Y., Shao, R., Wetzstein, G., Fei-Fei, L., Adeli, E.: The language of motion: Unifying verbal and non-verbal language of 3d human motion. In: arXiv (2024)
6. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: CVPR (2023)
7. Chen, Y., Zuo, R., Wei, F., Wu, Y., Liu, S., Mak, B.: Two-stream network for sign language recognition and translation. In: NeurIPS (2022)
8. Cheng, Q., Mayberry, R.I.: Acquiring a first language in adolescence: The case of basic word order in american sign language. Journal of child language **46**(2), 214–240 (2019)
9. Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., Joulin, A.: Beyond english-centric multilingual machine translation. In: arXiv (2020)
10. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: CVPR (2022)
11. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: MoMask: Generative masked modeling of 3d human motions. In: CVPR (2024)
12. Guo, C., Zuo, X., Wang, S., Cheng, L.: TM2T: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: ECCV (2022)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
14. Huang, W., Pan, W., Zhao, Z., Tian, Q.: Towards fast and high-quality sign language production. In: ACM MM (2021)
15. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV (2017)
16. Hwang, E.J., Kim, J.H., Cho, S.M., Park, J.C.: Non-autoregressive sign language production via knowledge distillation. In: BMVC (2021)
17. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: MotionGPT: Human motion as a foreign language. In: NeurIPS (2024)
18. Lee, T., Baradel, F., Lucas, T., Lee, K.M., Rogez, G.: T2LM: Long-term 3d human motion generation from multiple sentences. In: CVPR Workshop on Human Motion Generation (2024)
19. Lee, T., Moon, G., Lee, K.M.: MultiAct: Long-term 3d human motion generation from multiple action labels. In: AAAI (2023)
20. Li, S., Zhuang, S., Song, W., Zhang, X., Chen, H., Hao, A.: Sequential texts driven cohesive motions synthesis with natural transitions. In: ICCV (2023)
21. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out (2004)

22. 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)
23. 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)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2018)
25. Lu, S., Chen, L.H., Zeng, A., Lin, J., Zhang, R., Zhang, L., Shum, H.Y.: Human-tomato: Text-aligned whole-body motion generation. In: arXiv (2023)
26. Müller, M.: Information retrieval for music and motion, vol. 2. Springer (2007)
27. Napoli, D.J., Sutton-Spence, R.: Order of the major constituents in sign languages: Implications for all language. Frontiers in psychology **5**, 376 (2014)
28. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL (2002)
29. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. NeurIPS Workshop on Autodiff (2017)
30. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3D human motion synthesis with transformer VAE. In: ICCV (2021)
31. Petrovich, M., Black, M.J., Varol, G.: TEMOS: Generating diverse human motions from textual descriptions. In: ECCV (2022)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: arXiv (2021)
33. Saunders, B., Camgoz, N.C., Bowden, R.: Adversarial training for multi-channel sign language production. In: BMVC (2020)
34. Saunders, B., Camgoz, N.C., Bowden, R.: Progressive transformers for end-to-end sign language production. In: ECCV (2020)
35. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. In: ICLR (2024)
36. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: ICLR (2023)
37. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. In: NeurIPS (2017)
38. Wu, B., Xie, J., Shen, K., Kong, Z., Ren, J., Bai, R., Qu, R., Shen, L.: Mg-motionllm: A unified framework for motion comprehension and generation across multiple granularities. In: CVPR (2025)
39. 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)
40. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mT5: A massively multilingual pre-trained text-to-text transformer. In: NAACL (2021)
41. Yin, A., Li, H., Shen, K., Tang, S., Zhuang, Y.: T2S-GPT: Dynamic vector quantization for autoregressive sign language production from text. In: ACL (2024)
42. Yu, Z., Huang, S., Cheng, Y., Birdal, T.: SignAvatars: A large-scale 3d sign language holistic motion dataset and benchmark. In: ECCV (2024)
43. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2M-GPT: Generating human motion from textual descriptions with discrete representations. In: CVPR (2023)

44. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: MotionDiffuse: Text-driven human motion generation with diffusion model. IEEE TPAMI (2024)
45. Zhou, H., Zhou, W., Qi, W., Pu, J., Li, H.: Improving sign language translation with monolingual data by sign back-translation. In: CVPR (2021)
46. Zuo, R., Potamias, R.A., Ververas, E., Deng, J., Zafeiriou, S.: Signs as Tokens: A retrieval-enhanced multilingual sign language generator. In: arXiv (2024)
47. 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)