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

Xiao Liu[1†] Shiwei Gan[1†] Yafeng Yin[1*] Bowen Guo[1] Zhiwei Jiang[1]
Shunmei Meng[2] Lei Xie[1] Sanglu Lu[1]

[1] State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing 210023, China
[2] Department of Computer Science and Engineering, Nanjing University of Science and Technology, China

liuxiaox@smail.nju.edu.cn {sw,yafeng}@nju.edu.cn bowen@smail.nju.edu.cn
jzw@nju.edu.cn mengshunmei@njust.edu.cn {lxie,sanglu}@nju.edu.cn

Abstract

*Sign language production aims to generate sign sequences from spoken language, where the generation of sign pose sequences from text is often treated as a significant task. However, due to the differences in grammatical rules and modalities between sign language pose sequences and spoken language text, it is rather challenging to convert text into sign poses (*i.e.*, Text2Pose), while maintaining semantic consistency, motion accuracy and temporal coherence. In this paper, we focus on the Text2Pose task, and propose SignPR, a progressive diffusion framework that jointly models the structural and temporal properties of signing. Structurally, we perform progressive structural refinement: a structural VQVAE encodes each frame into semantic-aware and region-based discrete representations; the diffusion process first produces semantically consistent poses, and then progressively refines motion details under text and semantic conditioning. Temporally, we introduce block-wise causal diffusion, which progressively enforces temporal coherence and enables iterative refinement to earlier generated segments, yielding smoother transitions and reduced jitter. Extensive experiments on widely used datasets demonstrate that SignPR achieves superior results compared with prior T2P methods across multiple metrics, producing pose sequences that are semantically faithful, motion-accurate, and temporally coherent.*

1. Introduction

Sign Language Production (SLP) [6, 27, 31, 49] aims to convert spoken language into sign language (SL) sequences, thereby bridging communication barriers between deaf people and hearing people. Ideally, SL videos can effectively represent the intended sign language expressions. Therefore, there was some research work focusing on how to generate SL videos from texts [26]. However, due to the great difficulty in generating realistic and semantically consistent videos from text, recent studies have increasingly focused on generating sign pose sequences from text [6, 33]. In fact, converting text to sign poses is often treated as a significant task in SLP, since the generated sign poses can be further used for realistic video generation.

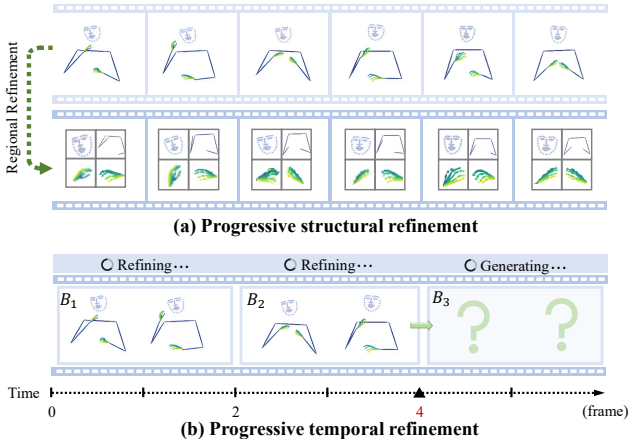

Figure 1. Our progressive SLP method.

When generating sign poses from text, some methods introduce glosses as intermediates to bridge grammatical gaps between sign and spoken language. Specifically, based on whether using glosses, prior work can be divided into Text2Gloss2Pose (T2G2P) and Text2Pose (T2P) categories. The former leverages gloss annotations to simplify alignment between the sign and spoken language [32, 33] , but glosses are scarce and costly to label [20, 43], limiting the practicality. The latter bypasses gloss supervision and maps text directly to continuous poses [31, 36]. Due to the lack of gloss annotations, T2P is usually a more challenging task.

[†]Equal contribution
[*]Yafeng Yin is the corresponding author.

In this paper, we focus on the T2P task, *i.e.*, generating sign poses from text without glosses. Due to the grammatical and modality differences between sign language and text, existing T2P methods often face challenges in semantic consistency, motion accuracy, and temporal coherence. **First, how to jointly consider semantic consistency and motion accuracy during modeling?** The structural modeling largely determines how well semantic intent and fine-grained motion details are preserved. To simplify pose modeling, many approaches employ a VQ-VAE to compress continuous pose features into discrete latent tokens. Broadly, these works fall into two categories: (i) Some frame-level models [23, 39, 44] compress and represent the whole body pose in one or multiple frames as a single token or ID, and adopt an autoregressive or diffusion-based model to predict the latent token sequence conditioned on text. However, such models mainly focus on semantics but lack sufficient capability to capture *motion accuracy*, since a single token provides limited capacity to represent fine-grained motion details. Yet, as a visual language, sign inherently depends on motion accuracy, reflected through accurate body movement, handshape, and facial expression. (ii) There are also models [42, 50] that focus on independently modeling local regions to enhance regional details. Although these approaches preserve local fidelity, they fall short on *global semantic consistency*. This is because limited inter-region interaction or simple fusion disrupts cross-region semantic and temporal alignment, yielding inconsistent compositions. **Second, how to achieve temporal coherence during generation?** Temporal modeling determines whether generated poses transition smoothly and causally over time. Existing T2P methods typically adopt either an autoregressive (AR) strategy [23, 44, 50] that predicts poses frame by frame, or a non-autoregressive strategy [1], such as diffusion, that predicts the entire sequence in parallel. AR-based models capture temporal causality well but suffer from exposure bias and slow decoding. While diffusion models offer diversity and fast parallel generation, they lack explicit temporal control. Especially in discrete diffusion, parallel temporal token updates further weaken causality, causing jitter and unsmooth inter-frame transitions.

Considering the above issues, we propose SignPR (as shown in Figure 1), a progressive diffusion framework that performs progressive refinement both structurally and temporally, jointly modeling the structural (semantic and motion) and temporal properties of sign language. Firstly, for semantic consistency and motion accuracy, we introduce progressive structural refinement. Specifically, we introduce a structural vector quantization variational autoencoder (VQVAE) that models pose at two different levels. At the semantic level, the VQVAE compresses and quantizes each frame-level pose into a single token unit to capture the whole-body structure, which reflects the semantic intent. At the motion level, it partitions the pose into key sub-regions (*e.g.*, body, hands, face) and quantizes each as a separate unit to emphasize fine-grained details, which improves motion accuracy. Built on these discrete representations, we propose a progressive diffusion model, which first generates semantically consistent poses, and then refines key region motion details conditioned on both the semantic-level pose tokens and the input text. Secondly, to ensure the temporal coherence of the generated pose sequence, we introduce progressive temporal refinement. Inspired by autoregressive causal ordering, we perform block-wise causal refinement instead of fully parallel diffusion, progressively enforcing temporal order and enabling iterative corrections to earlier generated segments. Our main contributions are summarized as follows:

- To ensure both semantic consistency and motion accuracy during modeling, we decouple pose generation into a structural, progressive modeling process. First, we design a structural VQVAE that models pose progressively with the semantic level and the motion level. Second, we introduce a structural diffusion model that first generates semantically consistent poses, and then refines local details under both semantic-level poses and text conditions.
- To ensure temporal coherence during the generation, we adopt a block-wise causal diffusion strategy, where poses are generated in temporal blocks and progressively refined to enforce causal consistency and mitigate temporal jitter.
- We propose SignPR, a progressive VQ-diffusion model that focuses on capturing the structural and temporal properties of signing during both modeling and generation. Extensive experiments on widely used datasets demonstrate that our model outperforms prior methods across evaluation metrics.

2. Related Work

Latent Diffusion Model. Latent diffusion models (LDMs) [3, 10] have shown impressive capabilities across generative tasks, including image synthesis [25, 28], video generation [41, 47], and motion generation [46]. During training, data are compressed into latent representations with an VAE; the LDM then adds noise to these latents and learns the corresponding noise distribution. After training, the model reverses this process: it starts from random noise, denoises the latent step by step, and the VAE decodes the final latent to a target sample. Early LDMs were typically designed for continuous latent spaces using VAEs. With the advent of VQ-VAE, VQ-Diffusion was proposed to model data distributions in a discrete latent space [8].

Sign Language Production. Most current SLP models [13, 14, 29, 40] have shifted attention to generating

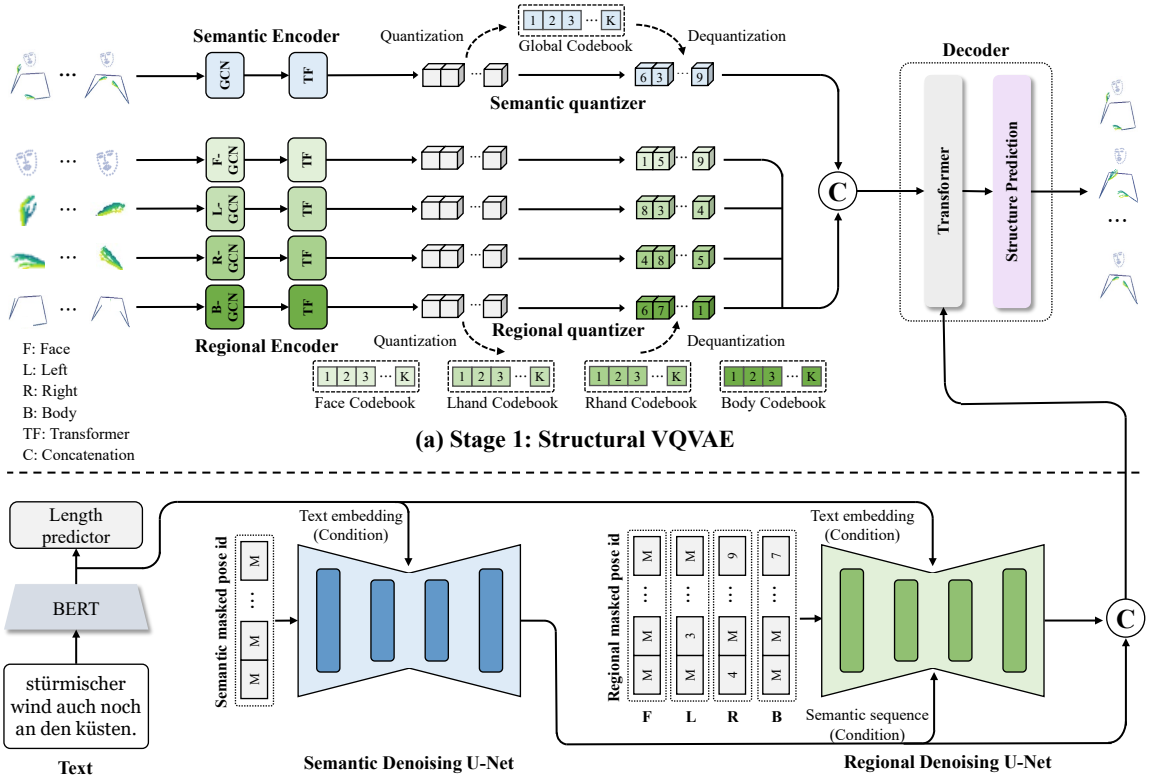

Figure 2. The progressive vector-quantized diffusion framework with S-VQVAE and S-Diffusion modules.

high-quality SL pose sequences, as it serves as an important condition for future SL video generation, when combined with off-the-shelf pose2video models [11]. As a result, the following pipelines have emerged, including Text2Gloss2Pose [32, 33, 38] , and Text2Pose [15, 31, 36] . Considering that obtaining gloss annotations is label-intensive, Text2Pose (T2P), which directly generates SL pose sequences from text without the need of gloss annotations, has gained more attention [1]. Early T2P models directly modeled sign sequences without VAE or VQ-VAE, which were not capable of generating high-quality pose sequences. Further, methods such as T2S-GPT [44], MS2SL [23] and SOKE [50] adopted VAE or VQ-VAE frameworks to encode pose sequences into a single continuous or discrete latent representation. Then, they adopted an AR model that generates pose sequences based on text and the previous predicted poses [16], or diffusion models that formulate pose generation as a denoising process (*i.e.*, progressively denoise noise into corresponding latent representations conditioned on text).

However, the motion accuracy in existing work usually remains limited. Specifically, many methods [23, 39, 44] treat the whole-body pose of one or multiple frames as a single token, where fine-grained hand and facial details are often lost, due to the limited capacity of a single unit. Some other methods [42, 50] generate different body regions as separate units, but this often results in uncoordinated body compositions or misaligned temporal relations. Besides, the existing work also suffers from the challenge of temporal coherence. Specifically, the autoregressive models [23, 44, 50] preserve temporal causality, but suffer from exposure bias. The diffusion models [1] often lack explicit temporal control, especially discrete diffusion with parallel token updates. These limitations motivate us to design a progressive discrete diffusion method that integrates both structural and temporal refinement, yielding accurate motions and coherent pose sequences. Unlike existing causal methods [9], our approach leaves diffusion training unchanged and improves temporal coherence through progressive block-wise refinement at inference time.

3. Method

Overall Framework. The goal of our T2P model is to generate a sequence of 2D SL poses $\mathbf{X} = \{x^s\}_{s=1}^{S}$, which not only semantically align with the corresponding text sequence c, but also exhibit natural and precise regional details. Here, $x^s \in \mathbb{R}^{J\times 2}$ denotes J pose keypoints in the s-th frame, and S denotes the number of frames. As shown in Figure 2, the process of our model can be simplified as follows. First, a Structural VQVAE (S-VQVAE) encoder quantizes the pose sequence $\mathbf{X}$ into two types of index

(id) sequences: an id sequence $I^{se} = \{i^{se,s}\}_{s=1}^{S}$ capturing high-level semantic information (*i.e.*, capturing whole-body dynamics rather than pose details), and four id sequences $I^{re} = \{i^{p,s}\}_{s=1}^{S}$ capturing details for each region $p \in \mathcal{P}$, where $\mathcal{P} = \{\text{body}, \text{right hand}, \text{left hand}, \text{head}\}$. As a result, the original pose sequence $\mathbf{X}$ can be represented by the combination of semantic and regional id sequences: I^{se} and I^{re}. Second, a lightweight two-layer Transformer is used to predict the target sequence length; and a Structural Diffusion Model (S-Diffusion) is adopted to progressively denoise random ids into the semantic id sequence $\hat{I}^{se}$ conditioned on the input text, which focuses on semantic consistency. Then, based on the generated semantic ids $\hat{I}^{se}$ and the input text, S-Diffusion further generates the regional id sequences $\hat{I}^{re}$, which capture fine-grained and part-wise specific pose details. Finally, after obtaining the semantic ids $\hat{I}^{se}$ and regional ids $\hat{I}^{re}$ from S-Diffusion, we employ the S-VQVAE decoder to reconstruct the final pose sequence $\hat{X}$ based on $\hat{I}^{se}$ and $\hat{I}^{re}$.

3.1. Stage 1: Pose Quantization with S-VQVAE

VQVAE transforms continuous SL pose generation into a discrete problem, reducing output dimensionality and helping reduce errors common in continuous regression. To obtain structured discrete representations of sign poses, we introduce a Structural VQVAE (S-VQVAE) that maps each frame to semantic latent ids for whole-body dynamics and regional latent ids for motion details.

Pose Quantization with Semantic VQVAE. We first propose a semantic VQVAE to obtain a semantic pose id sequence I^{se} for a sign pose sequence. The semantic VQVAE focuses on capturing the holistic/full-region structural features from the pose sequence, but may lack the capacity to model region-specific details of a pose. As illustrated in Figure 2, our semantic encoder $\mathcal{E}^{se}$ comprises a Graph Convolutional Network (GCN) [18] layer followed by two Transformer layers. The GCN layer is used to map a 2D pose sequence into pose embedding and the transformer layers are used to learn whole-body temporal relations among poses. Specifically, given a sequence of coordinates $\mathbf{X} = \{x^s\}_{s=1}^{S}$ of 2D pose, we first adopt the encoder to encode the joints as semantic latent embeddings $Z^{se} = \mathcal{E}^{se}(\mathbf{X})$, where $Z^{se} = \{z^{se,s}\}_{s=1}^{S}$. Then, we search the semantic codebook $\mathcal{C}^{\text{se}} = \{\mathcal{C}_j^{se}\}_{j=1}^{K^{se}}$, where K^{se} is the codebook size, to find the nearest neighbor i^s of $z^{se,s}$ based on the ℓ_2 distance in Eq. (1), aiming to get the discrete id sequence $I^{se} = \{i^s\}_{s=1}^{S}$ of the pose sequence.

$$i^{se,s} = \arg\min_{j} \|z^{se,s} - \mathcal{C}_j^{se}\|_2, \quad 1 \le j \le K^{se}; \quad \mathcal{C}_j^{se} \in \mathcal{C}^{se} \tag{1}$$

Pose Quantization with Regional VQVAE. To refine region pose details, we propose a regional VQVAE that captures region-specific features independently. Each sign pose is divided into four body-part regions: body, right hand, left hand, and head. For each region, a dedicated GCN layer encodes the spatial structure, followed by a region-specific two-layer Transformer that models temporal dependencies. This results in four streams of temporally encoded embeddings $Z^{re} = \{z^{p,s}\}_{p\in\mathcal{P},\, s=1}^{S}$, $\mathcal{P} = \{\text{body, right hand, left hand, head}\}$. Each region p is associated with a separate codebook $\mathcal{C}^p$, then we adopt the similar way in semantic VQVAE to get the nearest neighbor $i^{p,s}$ of $z^{p,s}$, and get the discrete id sequence $I^p = \{i^{p,s}\}_{s=1}^{S}$ of the pth region. In this way, each sign pose x^s is represented by four region-specific id sequences $\{I^p\}_{p\in\mathcal{P}}$.

During decoding, the quantized id sequences are mapped back to latent vectors, *i.e.*, get dequantized embeddings, via lookup from the corresponding codebooks. Then, the dequantized regional embeddings $\{\hat{z}^{p,s}\}$ and the semantic embedding $\hat{z}^{se,s}$ are concatenated, and decoded by the decoder $\mathcal{D}^{re}$ to reconstruct the pose sequence.

Structural Consistency Constraints for Semantic and Regional Latents. The latent spaces of semantic and regional VQVAEs should remain structurally consistent, since each whole-body semantic representation corresponds to a specific composition of regional details. To enforce this consistency, we introduce a structural consistency constraint that encourages the semantic latent to predict the regional latent ids of the same frame. Specifically, given the semantic latent feature $\hat{z}^{se,s}$ of frame s, a lightweight two-layer MLP ϕ^p is trained to predict the discrete regional id $i^{p,s}$ for each body part $p \in P$. The loss can be formulated as:

$$\mathcal{L}_{\text{cons}} = \frac{1}{S}\sum_{s=1}^{S}\sum_{p\in\mathcal{P}} \mathcal{L}_{\text{CE}}\left(\phi_p(\hat{z}^{se,s}),\, i^{p,s}\right) \tag{2}$$

Training of S-VQVAE. To train the proposed S-VQVAE, we combine four types of losses: an L1 reconstruction loss, a structural consistency loss shown in Eq. (2), a codebook commitment loss, and a quantization update loss.

$$\begin{aligned}\mathcal{L}_{\text{total}} = {} & \frac{1}{S}\sum_{s=1}^{S}\underbrace{\|\hat{x}^s - x^s\|_1}_{\text{Reconstruction}} + \underbrace{\lambda_{\text{cons}}\mathcal{L}_{\text{cons}}}_{\text{Consistency}} \\ & + \sum_{s=1}^{S}\underbrace{\beta^{se}\left\|\text{sg}[z^{se,s}] - \hat{z}^{se,s}\right\|_2^2}_{\text{Semantic commitment}} + \sum_{s=1}^{S}\underbrace{\beta^{re}\sum_{p\in\mathcal{P}}\left\|\text{sg}[z^{p,s}] - \hat{z}^{p,s}\right\|_2^2}_{\text{Regional commitment}} \\ & + \sum_{s=1}^{S}\underbrace{\gamma^{se}\left\|z^{se,s} - \text{sg}[\hat{z}^{se,s}]\right\|_2^2}_{\text{Semantic quantization}} + \sum_{s=1}^{S}\underbrace{\gamma^{re}\sum_{p\in\mathcal{P}}\left\|z^{p,s} - \text{sg}[\hat{z}^{p,s}]\right\|_2^2}_{\text{Regional quantization}}\end{aligned} \tag{3}$$

Here, $\hat{x}^s$ denotes the reconstructed pose for the sth pose. β^{se} and β^{re} control the balance of semantic and regional

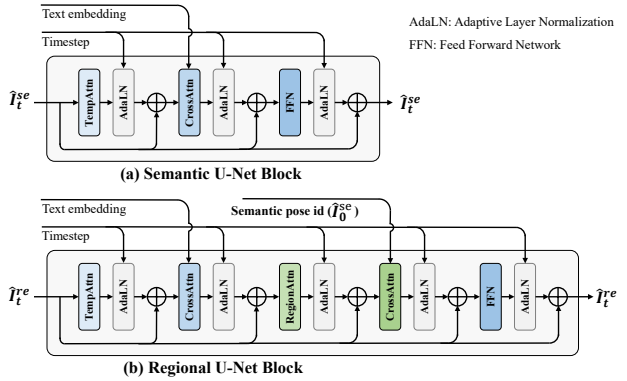

Figure 3. Semantic and regional denoising U-Nets.

branches. γ^{se} and γ^{re} control the balance of semantic and regional tokens. The stop-gradient operator sg$[\cdot]$ prevents gradients from flowing into the codebook during backpropagation. Regarding λ_{cons}, it controls the weight of the semantic consistency loss.

3.2. Stage 2: Pose Generation with S-Diffusion

Building upon the two-level latent space learned by S-VQVAE, we propose a Structural Discrete Diffusion Model (S-Diffusion), which generates sign pose sequences by using a diffusion model to progressively predict semantic and regional pose ids.

Semantic Discrete Diffusion Module. At the semantic level, the model learns a diffusion process from the input text c to the semantic pose id sequence I^{se}. In the training process, following VQ-Diffusion [8], we apply a discrete forward process $q(i_t^{se} \mid i_0^{se})$, where the ground-truth sequence $I_0^{se} \in \{1, \ldots, K^{se}\}^S$ is gradually corrupted via a Markov chain. Specifically, at each step, ids are independently replaced with a random codebook entry or a special `[MASK]` token, yielding the corrupted id sequence I_t^{se}.

In the reverse diffusion process, we follow G2P-DDM [42] and adopt a U-Net based denoising network ϕ_g to predict the denoised semantic pose id sequence $\hat{I}_0^{se}$ from noisy input c, where t^{se} is the diffusion timestep.

$$\hat{I}_0^{se} = \phi_g(I_t^{se}, t^{se}, c) \tag{4}$$

As shown in Figure 3(a), each semantic U-Net block comprises a temporal self-attention layer $\mathcal{TA}$ (*i.e.*, self attention is applied in temporal dimension), a cross-attention layer $\mathcal{CA}$ for the text c, and a feedforward module $\mathcal{FFN}$.

$$\begin{aligned} z_t^{se} &\leftarrow z_t^{se} + \mathcal{ALN}(\mathcal{TA}(z_t^{se}, t^{se})) \\ z_t^{se} &\leftarrow z_t^{se} + \mathcal{ALN}(\mathcal{CA}(z_t^{se}, t^{se}), c) \\ z_t^{se} &\leftarrow z_t^{se} + \mathcal{ALN}(\mathcal{FFN}(z_t^{se}, t^{se})) \end{aligned} \tag{5}$$

Here, $\mathcal{ALN}(\cdot)$ denotes adaptive layer normalization.

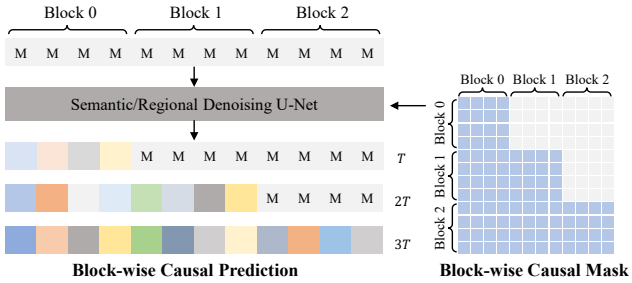

Figure 4. Block-wise causal inference process.

Regional Discrete Diffusion Module. To better generate subtle pose details, the regional diffusion performs region-specific denoising to generate four regional ids (*i.e.*, body, left hand, right hand, and head), while conditioned on the input text c. Besides, to ensure that the regional diffusion focuses more on pose details while maintaining semantic consistency, we further incorporate semantic ids $\hat{I}_0^{se}$ as conditions during the denoising process, as shown below.

$$\hat{I}_0^{re} = \phi_l(I_t^{re}, t^{re}, c, \hat{I}_0^{se}) \tag{6}$$

Here, ϕ_l is a U-Net based denoising network, $\hat{I}_0^{re} = \{\hat{I}_0^p\}_{p \in \mathcal{P}}$ denotes the set of region-wise id sequences, I_t^{re} is the corresponding noisy input, and t^{re} is the timestep. The regional U-Net retains the semantic backbone but adds following two modules: a region attention $\mathcal{RT}$ to facilitate information exchange across regions, and an extra cross-attention $\mathcal{CA}_{se}$ conditioned on the semantic pose:

$$\begin{aligned} z_t^p &\leftarrow z_t^p + \mathcal{ALN}(\mathcal{RT}(z_t^p), t^{re}) \\ z_t^p &\leftarrow z_t^p + \mathcal{ALN}(\mathcal{CA}_{se}(z_t^p, \hat{I}_0^{se}), t^{re}) \end{aligned} \tag{7}$$

3.3. Causal Inference via Progressive Refinement

Normally, discrete diffusion updates all time tokens in parallel at each denoising step, which weakens temporal causality due to the absence of explicit causal modeling, leading to jittery sequences. To address this issue, we propose a novel inference paradigm named causal inference via progressive refinement (InferRef), which requires no retraining and improves the temporal smoothness and overall quality of the generated SL videos.

Specifically, inspired by autoregressive decoding, InferRef adopts a block-wise causal strategy that progressively refines the generated SL sequence. Before inference, we use the length predictor in Figure 2 to estimate the target frame length S, *i.e.*, determining the number of tokens to be denoised. As shown in Figure 4, given a length-S discrete token sequence $I = \{i_s\}_{s=1}^S$ at an inference step, we fix a block size K and partition I into $N = \lceil S/K \rceil$ contiguous blocks $I = \{B_i\}_{i=0}^{N-1}$. The generation is modeled as

$$p(I \mid c) = \prod_{i=1}^{N-1} p\big(B_i \mid c, B_{<i}\big), \tag{8}$$

where c is the text condition and $B_{<i} = \{B_j\}_{j=0}^{i-1}$. We apply a block-wise causal mask that allows tokens in B_i to attend to all tokens in $B_{<i}$ and intra-block tokens in B_i, while masking future blocks. When predicting B_i, previously generated blocks $B_{<i}$ are refined simultaneously, yielding temporally coherent, smooth SL sequences.

After applying InferRef in both the semantic and regional diffusion inference process, we obtain the final denoised token sequences $\hat{I}_0^{se}$ and $\hat{I}_0^{re}$, corresponding to the semantic and region-specific pose ids, respectively. These are then mapped back to get semantic embedding $\hat{Z}^{se}$ and regional embeddings $\hat{Z}^{p}$ via codebook lookup. Then, the final pose is reconstructed by decoding the concatenation of $\hat{Z}^{se}$ and $\hat{Z}^{p}$ using S-VQVAE decoder.

4. Experiments

Datasets. We evaluate our model on three publicly available SL datasets. (1) **Phoenix14T** [4] is a German SL dataset with 7096 training, 519 validation, and 642 test samples from 9 signers. It contains both gloss and translation annotations with a vocabulary of 1066 glosses and 2877 German words. (2) **CSL-Daily** [48] is a Chinese SL dataset containing 18401, 1077, and 1176 videos for training, validation, and testing from 10 signers. It provides both gloss and translation annotations with 2000 glosses and 2343 words. (3) **USTC-CSL** [12] is a Chinese SL dataset with 20000 and 5000 SL videos for training and testing from 50 signers, covering 3100 words in translation annotations.

Evaluation Metrics. Following previous T2P and G2P work [42, 44] , we adopt a SL translation model [7] to perform back-translation, which translates the generated pose sequences back into text. We report ROUGE-L [21] and BLEU-1 to BLEU-4 scores [24] to evaluate the semantic accuracy of generated sign sequences. Additionally, we apply Dynamic Time Warping Mean Joint Error (DTW-MJE) [2] to measure temporal alignment between generated and ground-truth pose sequences. To further assess motion accuracy, we compute Fréchet Inception Distance (FID) and Mean Per Joint Position Error (MPJPE). Specifically, MPJPE is measured on keypoint coordinates, while FID is computed on pose features extracted by the frozen encoder of our trained S-VQVAE.

Implementation Details. For pose extraction, we use HRNet [45] to estimate 2D keypoints for each frame, including 42 hand keypoints, 68 facial keypoints, and 11 upper-body keypoints. For S-VQVAE, the codebook embedding dimension is 768, and both the semantic and regional codebooks contain 1024 entries. S-VQVAE and S-Diffusion comprise 75.50M and 680.04M parameters, respectively. Both S-VQVAE and S-Diffusion are trained with the AdamW optimizer [22], using a batch size of 32 for 100K iterations. The initial learning rate is set to 1×10^{-4} and follows a cosine decay schedule. All experiments are implemented in PyTorch and conducted on 8 NVIDIA A6000 GPUs.

4.1. Comparisons

*For a fair comparison, we mainly reproduce baselines with publicly available resources (*i.e.*, PT, G2P-DDM, Sign-IDD) and report our reproduced back-translation results under the same experimental settings.*

Evaluation on Phoenix14T Dataset. As shown in Table 1, we compare our method with both T2G2P models and T2P models. Usually, the T2G2P models yield stronger results than previous T2P models, highlighting the benefit of introducing glosses as an intermediate representation. However, despite lacking gloss supervision, our SignPR achieves the best performance across all metrics. Specifically, compared with the SOTA T2G2P model Sign-IDD, SignPR improves BLEU-1 from 26.45 to 31.91 and reduces MPJPE from 47.19 to 23.04. Compared with the T2P model MoMP, SignPR achieves a BLEU-1 score of 31.91 (vs. 16.87) and a lower FID of 2.15 (vs. 2.97).

Evaluation on CSL-Daily Dataset. Table 2 shows that SignPR achieves clear improvements over the baseline MoMP. Specifically, it improves BLEU-4 from 2.14 to 3.01 (+0.87). In addition, it reduces FID from 2.95 to 2.47 (-0.48) and MPJPE from 48.24 to 44.22 (-4.02), indicating enhanced generated pose quality.

Evaluation on USTC-CSL Dataset. We evaluate on USTC-CSL under two data splits: Split-I and Split-II. As shown in Table 3, our method consistently outperforms PT-GN across all metrics. On Split-I, it improves ROUGE from 25.09 to 95.48 and reduces MPJPE from 183.14 to 51.40. Split-II shows a similar trend, confirming clear gains in both semantic quality and motion accuracy.

4.2. Qualitative Results

Visualization Comparison. We visualize sign pose samples from the ground truth (GT), MoMP, SignPR w/o regional, and SignPR. Back translation is used to assess the semantic accuracy of the generated poses. As shown in Figure 7, MoMP fails to reflect the correct textual semantics and produces multiple incorrect poses (red). The SignPR w/o regional variant captures the coarse structure but lacks accurate pose details, leading to regional errors (yellow) and semantic drift. In contrast, SignPR generates accurate, coherent poses, and the back-translation recovers a sentence that is mostly correct, with few regional errors.

Text2Gloss2Pose	DEV						TEST					
	ROUGE	BLEU1	BLEU4	FID↓	MPJPE↓	DTW-MJE↓	ROUGE	BLEU1	BLEU4	FID↓	MPJPE↓	DTW-MJE↓
PT-base [30]	8.61	9.53	0.72	2.90	41.92	0.175	8.88	9.47	0.59	3.22	51.35	0.206
PT-GN [30]	11.87	12.51	3.88	2.98	40.63	0.171	13.17	13.35	4.31	3.33	50.80	0.204
G2P-DDM [42]	20.05	19.97	6.36	2.65	39.34	0.119	20.37	19.40	6.25	2.91	40.02	0.116
Sign-IDD [37]	27.97	26.76	8.42	2.22	39.11	0.116	27.11	26.45	8.66	2.46	47.19	0.118
Text2Pose	DEV						TEST					
PT [30]	8.36	6.11	0.05	3.32	48.60	0.213	8.44	6.22	0.01	3.41	49.37	0.213
PT-GN [30]	8.38	7.18	0.01	3.32	48.21	0.212	8.57	9.22	0.01	3.39	48.30	0.211
MoMP [32]	16.16	16.30	4.21	3.15	40.96	0.172	17.02	16.87	4.58	2.97	45.71	0.168
SignPR	**30.84**	**31.06**	**9.12**	**2.13**	**23.77**	**0.114**	**32.86**	**31.91**	**9.41**	**2.15**	**23.04**	**0.111**

Table 1. Comparison of SLP performance on the PHOENIX-14T dataset.

Text2Pose	TEST					
	ROUGE	BLEU1	BLEU4	FID↓	MPJPE↓	DTW-MJE↓
PT-GN [30]	7.54	7.12	0.41	3.58	90.28	0.428
MoMP [32]	11.67	11.62	2.14	2.95	48.24	0.129
SignPR	**14.43**	**14.08**	**3.01**	**2.47**	**44.22**	**0.108**

Table 2. Comparison of SLP performance on the CSL-Daily.

Text2Pose	TEST (Split-I)			TEST (Split-II)		
	ROUGE	BLEU-4	MPJPE↓	ROUGE	BLEU-4	MPJPE↓
PT-GN [30]	25.09	10.65	183.14	15.22	3.95	237.63
SignPR	**95.48**	**94.54**	**51.40**	**42.10**	**17.16**	**64.92**

Table 3. Comparison of SLP performance on the USTC-CSL.

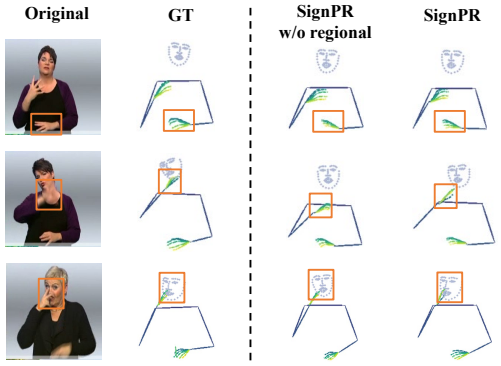

Figure 5. Visualization results with/without regional refinement.

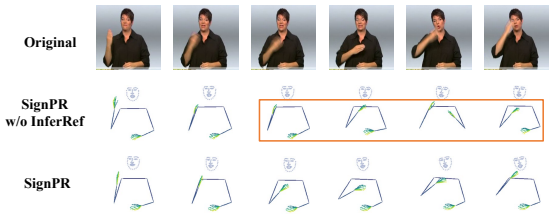

Figure 6. Visualization results with/without temporal refinement.

Visualization of Regional Details and Temporal Coherence. As shown in Figure 5, SignPR produces more accurate regional pose details than the variant without regional refinement, especially in finger shapes (row 1). Beyond enhancing details, the regional module also compensates for semantic stage errors, correcting arm placement (row 2) and improving facial expressions (row 3), leading to more faithful and natural sign generation. As shown in Figure 6, when causal inference is removed (row 2), the generated sequences of SignPR exhibit clear motion discontinuities and frame-to-frame jitter, along with incorrect poses. In contrast, with causal inference (row 3), these issues are greatly alleviated, demonstrating its effectiveness in improving temporal coherence.

4.3. Ablation Study on S-VQVAE.

We perform all ablation studies on the Phoenix14T dataset.

Effect of S-VQVAE on Reconstruction Quality. Table 4 evaluates different codebook configurations on pose reconstruction (not generation from text). Regional codebooks improve rBLEU, rROUGE and reduce rMPJPE and rDTW, showing better capture of fine-grained pose details. Combining semantic and regional codebooks yields further gains, indicating that modeling both semantics and regional structure improves reconstruction quality.

Effect of Key Components in S-VQVAE. As shown in Table 5, replacing the GCN with a Transformer causes a clear performance drop, highlighting the advantage of GCN for modeling human poses as structured graphs. Removing the structural consistency loss $\mathcal{L}_{\text{cons}}$ further hurts performance, indicating that enforcing structural alignment between semantic and regional representations is crucial.

4.4. Ablation Study on S-Diffusion.

Effect of S-Diffusion on Generation Quality. Table 6 reports the generation performance under different diffusion configurations. The variants using semantic diffusion only or regional diffusion only show inferior performance. While using the structural diffusion framework significantly improves generation quality, with notable gains in all metrics. Combining semantic and regional diffusion yields the strongest results.

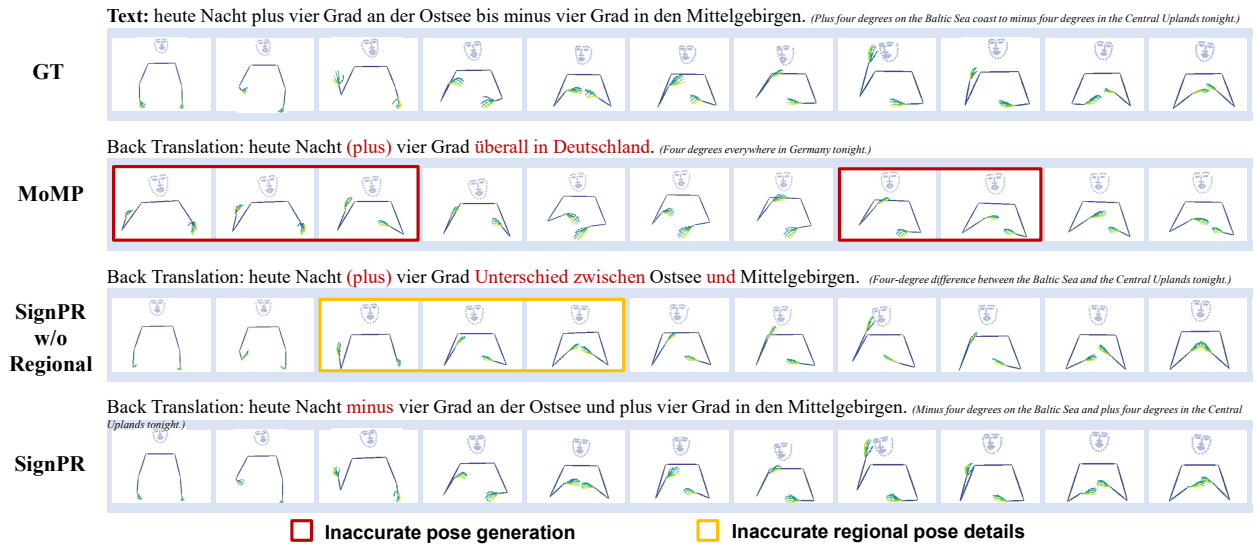

Figure 7. Qualitative comparison of generated poses from SignPR, the variant (w/o regional) of SignPR and MoMP.

Configuration		VQVAE Reconstruction					
Semantic-codebook	Regional-codebook	rROUGE	rBLEU1	rBLEU4	rFID↓	rMPJPE↓	rDTW-MJE↓
✓		36.69	36.30	10.73	0.76	2.8	0.002
	✓	37.39	37.13	11.30	0.27	2.6	0.001
✓	✓*	37.54	37.06	11.24	0.25	2.5	0.001
✓	✓	**39.51**	**39.65**	**12.92**	**0.19**	**2.3**	**0.001**

Table 4. Effect of structural VQVAE on reconstruction quality. *Note:* Metrics with prefix 'r' reflect VQVAE reconstruction quality. '✓*' indicates shared regional codebook across regions.

Variant	ROUGE	BLEU1	BLEU4	FID↓	MPJPE↓	DTW-MJE↓
w/o $\mathcal{L}_{\text{cons}}$	24.35	24.08	7.81	2.80	38.63	0.121
w/o GCN	26.64	26.65	7.70	2.35	28.07	0.115
SignPR	**32.86**	**31.91**	**9.41**	**2.15**	**23.04**	**0.111**

Table 5. Effect of components in proposed S-VQVAE.

Semantic-Dif	Regional-Dif	Diffusion Generation					
		ROUGE	BLEU1	BLEU4	FID↓	MPJPE↓	DTW-MJE↓
✓		22.53	22.26	6.31	2.85	42.96	0.124
	✓	23.90	23.12	7.53	2.58	40.32	0.122
✓	✓*	27.49	27.26	8.15	2.63	26.06	0.115
✓	✓	**32.86**	**31.91**	**9.41**	**2.15**	**23.04**	**0.111**

Table 6. Effect of structural diffusion on generation quality. *Note:* ✓* indicates using shared regional codebook across regions.

Effect of Regional Diffusion Conditioning. Table 7 shows that conditioning on text or semantic pose alone is insufficient; both semantic and text are needed. The integration method also matters: simple concatenation yields moderate gains, whereas the cross-attention (CA) used in SignPR performs best.

Effect of Block-wise Causal Inference. As shown in Table 8, we vary the frames per block K in block-wise causal inference. Using $K = \infty$ (no blocking) as the base, smaller K yields more blocks and increases inference time. At the same time, quality consistently improves. Under the current experimental setup, using K=8 yields better results, with ROUGE improved by +1.36 over the base.

Variant	ROUGE	BLEU1	BLEU4	FID↓	MPJPE↓	DTW-MJE↓
Text only	23.90	23.12	7.53	2.58	40.32	0.122
Semantic Pose only	24.72	24.48	7.69	2.57	36.11	0.118
Text + Semantic Pose (Concat)	26.42	26.21	8.72	2.44	30.10	0.115
Text + Semantic Pose (CA)	**32.86**	**31.91**	**9.41**	**2.15**	**23.04**	**0.111**

Table 7. Effect of regional diffusion conditioning.

frames/block	s/video	ROUGE	BLEU1	BLEU4	FID↓	MPJPE↓	DTW-MJE↓
∞	1.47	31.50	31.24	9.16	2.20	23.79	0.114
32	1.95	32.03	31.62	9.25	2.28	23.40	0.112
16	2.11	32.27	31.80	9.30	2.17	23.23	0.112
8	2.45	**32.86**	**31.91**	**9.41**	**2.15**	**23.04**	**0.111**

Table 8. Ablation on the block-wise causal inference.

5. Conclusion

In this work, we focus on the T2P task, aiming to generate semantically consistent, motion-accurate, and temporally coherent sign pose sequences from text. We introduced SignPR, a progressive diffusion framework that jointly models the structural and temporal properties of signing. Through structural progressive refinement with semantic and regional generation and temporal progressive refinement via block-wise causal inference, SignPR achieves superior semantic consistency, motion accuracy, and temporal coherence. Extensive experiments on three datasets show that our method achieves superior performance.

6. Acknowledgments

This work is supported in part by Key Projects of Jiangsu Provincial Basic Research Program under Grant No. BK20243040; National Natural Science Foundation of China under Grant Nos. 62172208, 92467202, 62272216; JiangSu Natural Science Foundation under Grant No. BK20251989. This work is partially supported by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM118); Collaborative Innovation Center of Novel Software Technology and Industrialization. This work is also supported by the 2025 Jiangsu Frontier Technology R&D Project (No. BF2025004) on Trusted Data mining for Cross-Domain Data Security Sharing and Privacy Protection.

References

[1] Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. Neural sign actors: A diffusion model for 3d sign language production from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1985–1995, 2024. 2, 3

[2] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd international conference on knowledge discovery and data mining*, pages 359–370, 1994. 6

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[4] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793, 2018. 6

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[6] Sen Fang, Chunyu Sui, Xuedong Zhang, and Yapeng Tian. Signdiff: Learning diffusion models for american sign language production. *arXiv e-prints*, pages arXiv–2308, 2023. 1

[7] Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Lei Xie, Sanglu Lu, and Hongkai Wen. Mixsigngraph: A sign sequence is worth mixed graphs of nodes. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 6

[8] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10696–10706, 2022. 2, 5

[9] Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. Ssd-lm: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11575–11596, 2023. 3

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[11] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 3

[12] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. Video-based sign language recognition without temporal segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 6

[13] Wencan Huang, Wenwen Pan, Zhou Zhao, and Qi Tian. Towards fast and high-quality sign language production. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3172–3181, 2021. 2

[14] Wencan Huang, Zhou Zhao, Jinzheng He, and Mingmin Zhang. Dualsign: semi-supervised sign language production with balanced multi-modal multi-task dual transformation. In *Proceedings of the 30th ACM international conference on multimedia*, pages 5486–5495, 2022. 2

[15] Eui Jun Hwang, Jung-Ho Kim, and Jong C Park. Non-autoregressive sign language production with gaussian space. In *BMVC*, page 197, 2021. 3

[16] Eui Jun Hwang, Huije Lee, and Jong C Park. Autoregressive sign language production: A gloss-free approach with discrete representations. *arXiv preprint arXiv:2309.12179*, 2023. 3

[17] Hyeonho Jeong, Gihyun Kwon, and Jong Chul Ye. Zero-shot generation of coherent storybook from plain text story using diffusion models. *arXiv preprint arXiv:2302.03900*, 2023.

[18] TN Kipf. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 4

[19] Dimitris Kouremenos, Klimis S Ntalianis, Giorgos Siolas, and Andreas Stafylopatis. Statistical machine translation for greek to greek sign language using parallel corpora produced via rule-based machine translation. In *CIMA@ ICTAI*, pages 28–42, 2018.

[20] Shyam Krishna and Janmesh Ukey. Gan based indian sign language synthesis. In *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–8, 2021. 1

[21] Chin-Yew Lin and Franz Josef Och. 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)*, pages 605–612, 2004. 6

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[23] Jian Ma, Wenguan Wang, Yi Yang, and Feng Zheng. Ms2sl: multimodal spoken data-driven continuous sign language production. *arXiv preprint arXiv:2407.12842*, 2024. 2, 3

[24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[25] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI workshop on deep generative models*, pages 117–126. Springer, 2022. 2

[26] Fan Qi, Yu Duan, Huaiwen Zhang, and Changsheng Xu. Signgen: End-to-end sign language video generation with latent diffusion. In *European Conference on Computer Vision*, pages 252–270. Springer, 2024. 1

[27] Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and Mohammad Sabokrou. Sign language production: A review. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3451–3461, 2021. 1

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[29] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Adversarial training for multi-channel sign language production. *arXiv preprint arXiv:2008.12405*, 2020. 2

[30] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Progressive transformers for end-to-end sign language production. In *European Conference on Computer Vision*, pages 687–705. Springer, 2020. 7

[31] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks. *International journal of computer vision*, 129(7):2113–2135, 2021. 1, 3

[32] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Mixed signals: Sign language production via a mixture of motion primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1919–1929, 2021. 1, 3, 7

[33] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5141–5151, 2022. 1, 3

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[35] Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. Sign language production using neural machine translation and generative adversarial networks. In *Proceedings of the 29th British Machine Vision Conference (BMVC 2018)*. British Machine Vision Association, 2018.

[36] Stephanie Stoll, Armin Mustafa, and Jean-Yves Guillemaut. There and back again: 3d sign language generation from text using back-translation. In *2022 International Conference on 3D Vision (3DV)*, pages 187–196. IEEE, 2022. 1, 3

[37] Shengeng Tang, Jiayi He, Dan Guo, Yanyan Wei, Feng Li, and Richang Hong. Sign-idd: Iconicity disentangled diffusion for sign language production. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7266–7274, 2025. 7

[38] Shengeng Tang, Feng Xue, Jingjing Wu, Shuo Wang, and Richang Hong. Gloss-driven conditional diffusion models for sign language production. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21 (4):1–17, 2025. 3

[39] Harry Walsh, Abolfazl Ravanshad, Mariam Rahmani, and Richard Bowden. A data-driven representation for sign language production. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10. IEEE, 2024. 2, 3

[40] Xu Wang, Shengeng Tang, Peipei Song, Shuo Wang, Dan Guo, and Richang Hong. Linguistics-vision monotonic consistent network for sign language production. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 2

[41] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 133(5):3059–3078, 2025. 2

[42] Pan Xie, Qipeng Zhang, Peng Taiying, Hao Tang, Yao Du, and Zexian Li. G2p-ddm: Generating sign pose sequence from gloss sequence with discrete diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6234–6242, 2024. 2, 3, 5, 6, 7

[43] Aoxiong Yin, Tianyun Zhong, Li Tang, Weike Jin, Tao Jin, and Zhou Zhao. Gloss attention for gloss-free sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2551–2562, 2023. 1

[44] Aoxiong Yin, Haoyuan Li, Kai Shen, Siliang Tang, and Yueting Zhuang. T2s-gpt: Dynamic vector quantization for autoregressive sign language production from text. *arXiv preprint arXiv:2406.07119*, 2024. 2, 3, 6

[45] Changqian Yu, Bin Xiao, Changxin Gao, Lu Yuan, Lei Zhang, Nong Sang, and Jingdong Wang. Lite-hrnet: A lightweight high-resolution network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10440–10450, 2021. 6

[46] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 2

[47] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 2

[48] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with

monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1316–1325, 2021. 6

[49] Ronglai Zuo, Fangyun Wei, Zenggui Chen, Brian Mak, Jiaolong Yang, and Xin Tong. A simple baseline for spoken language to sign language translation with 3d avatars. In *European Conference on Computer Vision*, pages 36–54. Springer, 2024. 1

[50] Ronglai Zuo, Rolandos Alexandros Potamias, Evangelos Ververas, Jiankang Deng, and Stefanos Zafeiriou. Signs as tokens: A retrieval-enhanced multilingual sign language generator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23806–23816, 2025. 2, 3