# Hybrid Autoregressive-Diffusion Model for Real-Time Sign Language Production

**Maoxiao Ye  Xinfeng Ye  Mano Manoharan**
University of Auckland, New Zealand
{xye804@aucklanduni.ac.nz, x.ye@auckland.ac.nz, mano.manoharan@auckland.ac.nz}

## Abstract

Earlier Sign Language Production (SLP) models typically relied on autoregressive decoding, which naturally preserves temporal causality but suffers from error accumulation at inference time. More recent diffusion-based approaches improve generation quality through iterative denoising, yet their sequence-level refinement process introduces substantial latency. To address this trade-off, we propose **HybridSign**, a hybrid autoregressive-diffusion model for low-latency sign language production that combines causal frame generation with flow-based diffusion refinement. A **Multi-Scale Pose Representation** module captures fine-grained articulator features, while a **Confidence-Aware Causal Attention** mechanism leverages joint-level confidence scores to improve robustness under noisy 2D pose observations. Experiments on PHOENIX14T and How2Sign show that HybridSign consistently achieves the best quality–efficiency trade-off among the compared baselines. On the How2Sign test split, it reaches BLEU-1/4 scores of 30.12/6.48 and DTW of 3.89, while reducing time-to-first-frame to 5.90 s and increasing throughput to 10.17 FPS under a 60-frame evaluation protocol.

## 1 Introduction

Sign language relies on coordinated body, hand, and facial motion, making automatic sign language production a challenging structured generation problem. A practical SLP system must preserve temporal coherence, articulate fine-grained local motion, and respond with low latency. Existing approaches typically favor either autoregressive models (Saunders et al., 2020a,b; Tang et al., 2022), which model temporal dependencies well but suffer from exposure bias during inference, or diffusion models (Huang et al., 2021; Xie et al., 2024; Tang et al., 2025), which generate higher-quality poses but incur substantial sampling cost.

We propose **HybridSign**, a hybrid autoregressive-diffusion model that combines the strengths of both paradigms for low-latency SLP. In this paper, low latency refers to the ability to emit the first pose frame quickly and then continue generation frame by frame, which is more informative for interactive use than reporting only the total offline generation time. The autoregressive pathway provides causal frame generation, while the flow-based diffusion pathway improves per-frame refinement quality. A **Multi-Scale Pose Representation** design captures complementary face, body, and hand dynamics, and a **Confidence-Aware Causal Attention** mechanism uses keypoint reliability to improve robustness. Our main contributions are summarized as follows:

- A hybrid autoregressive-diffusion framework for low-latency SLP that combines frame-wise causal generation with flow-based diffusion refinement.
- A three-expert multi-scale pose representation and fusion design that preserves fine-grained articulator details while explicitly modeling coupled bimanual motion.
- A confidence-aware causal attention mechanism and a self-forcing training protocol for improving robustness under noisy 2D pose observations.

## 2 Related Work

Sign language is a primary means of communication for deaf and hard-of-hearing communities, yet communication barriers remain common because many hearing individuals do not know sign language. Early research therefore focused mainly on Sign Language Recognition (SLR) (Guo et al., 2017; Hu et al., 2023a,b; Koller, 2020; Zhao et al.,



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 750–763

2023) and Sign Language Translation (SLT) (Camgoz et al., 2020a; Gong et al., 2024; Guo et al., 2019). More recently, increasing attention has been directed toward SLP, which aims to generate sign motion from linguistic input.

### 2.1 Sign Language Production

Initial studies in SLP predominantly relied on rule-based animation techniques to translate textual inputs into synthetic avatar-based sign language animations (Mazumder et al., 2021; McDonald et al., 2016; Segouat, 2009). These methods typically employed predefined templates and handcrafted lookup rules to map sentences or glosses to gestures. While these systems offered interpretable sign animations, they suffered from high data collection costs, limited scalability, and poor generalization to unseen sentences or grammatical structures.

### 2.2 Autoregressive models in SLP

With the advent of deep learning, more flexible and data-driven SLP models were introduced. Text2Sign (Stoll et al., 2020) introduces a multi-stage pipeline for SLP, which decomposes the overall process into three sequential subtasks: text-to-gloss[1] translation, gloss-to-pose generation, and pose-to-video synthesis. While effective, such pipeline designs introduced potential efficiency issues and error propagation between stages. To address this, (Saunders et al., 2020b) presented the first end-to-end autoregressive model that directly generates sign pose sequences from glosses. Further improvements followed, including a multi-channel adversarial model (Saunders et al., 2020a) and a Mixture Density Transformer (Saunders et al., 2021) to better model the multi-modal nature of sign gestures.

### 2.3 Diffusion models in SLP

Recently, diffusion models have shown strong potential in SLP due to their capacity for modeling complex output distributions. Methods such as SignDiff (Fang et al., 2025), G2P-DDM (Xie et al., 2024), and GCDM (Tang et al., 2025) leverage denoising diffusion probabilistic models (DDPMs (Ho et al., 2020)) to generate sign pose sequences by gradually refining Gaussian noise into meaningful joint coordinates. These approaches typically treat pose generation as a coordinate regression problem, guided by gloss or semantic input. Subsequently, Sign-IDD (Tang et al., 2024b) integrates an attribute-controllable diffusion module that leverages skeletal direction and length attributes to constrain joint associations, thereby enabling more precise and controllable pose generation.

[1]Gloss is a written representation of signs using capitalized words to show the meaning and order of the signs.

## 3 Preliminary

In this section, we introduce the fundamental concepts and notations used in our framework, including autoregressive modeling, diffusion-based generation, human pose representation, and attention mechanisms.

### 3.1 Autoregressive Modeling

Autoregressive models generate a sequence by modeling each element conditioned on its previous elements. For a sequence of human poses $P = \{P_1, P_2, \ldots, P_T\}$, where $P_t \in \mathbb{R}^{J \times D}$ represents the positions of $J$ joints at time $t$ in $D$-dimensional space, the autoregressive factorization is:

$$p(P) = \prod_{t=1}^{T} p(P_t \mid P_{<t}) \tag{1}$$

This formulation allows for temporal modeling but is prone to error accumulation due to its greedy generation process.

### 3.2 Diffusion Models

**Score-based diffusion models.** Diffusion models learn to generate data by reversing a gradual noising process. Traditional diffusion models (Song et al., 2021, 2020; Nichol and Dhariwal, 2021) follow the denoising diffusion probabilistic model (DDPM) framework (Ho et al., 2020), where a data sample $x_0$ is progressively perturbed over $T$ steps using a predefined noise schedule $\{\beta_t\}_{t=1}^{T}$. The forward process is defined as:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{2}$$

This results in:

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t) I) \tag{3}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

The denoising model is trained to predict the added noise $\boldsymbol{\epsilon}$:

$$\mathcal{L}_{\text{Diff}} = \mathbb{E}_{x_0,t,\boldsymbol{\epsilon}} \left[ \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(x_t, t)\|^2 \right] \quad (4)$$

However, traditional diffusion models have slow sampling speeds, which significantly hinder low-latency generation.

**Flow-based Diffusion Models.** In contrast to DDPMs, flow-based diffusion models (Lipman et al., 2024, 2023; Liu et al., 2022) learn an explicit transformation between noise and data using invertible mappings. These models parameterize the denoising process as an ODE (or deterministic flow) rather than a stochastic process. Specifically, they model a continuous-time trajectory between the data distribution and a base noise distribution using a neural network as the vector field.

The process is defined through the following differential equation:

$$\frac{dx(t)}{dt} = v_\theta(x(t), t), \quad x(0) \sim \mathcal{N}(0, I) \quad (5)$$

Here, $v_\theta$ is a neural network trained to match the score or displacement between perturbed data and clean data. This flow transports samples from the base distribution (e.g., Gaussian) at $t = 0$ to the data distribution at $t = 1$.

A popular training objective is the flow matching loss:

$$\mathcal{L}_{\text{Flow}} = \mathbb{E}_{x_0,t,z} \left[ \left\| v_\theta(x_t, t) - \frac{x_0 - x_t}{t} \right\|_2^2 \right] \quad (6)$$

where $x_t = (1-t)z + tx_0$ is a linear interpolation between a noise sample $z \sim \mathcal{N}(0, I)$ and a data sample $x_0$.

This objective encourages the model to learn a displacement field that transports noise into data along straight paths, offering a deterministic alternative to stochastic diffusion, thereby accelerating training efficiency and sampling speed.

### 3.3 Human Pose Representation

A human pose at time $t$ is represented as a set of $J$ joints:

$$P_t = \{p_{t,1}, p_{t,2}, \ldots, p_{t,J}\}, \quad p_{t,j} \in \mathbb{R}^D \quad (7)$$

Each joint is also associated with a confidence score $c_{t,j} \in [0, 1]$, denoted as:

$$C_t = \{c_{t,1}, c_{t,2}, \ldots, c_{t,J}\} \quad (8)$$

These confidence scores are used later in our attention mechanism to modulate the influence of each joint during pose generation (Section 4.4).

### 3.4 Attention Mechanism

We build upon the standard scaled dot-product attention mechanism. Given query $Q \in \mathbb{R}^{n \times d}$, key $K \in \mathbb{R}^{m \times d}$, and value $V \in \mathbb{R}^{m \times d}$ matrices, the attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) V \quad (9)$$

To preserve temporal causality in generation, we apply a causal mask to prevent attention to future time steps.

Later, in our proposed method, we extend this formulation to include confidence-awareness, biasing the attention distribution based on joint reliability (Section 4.4).

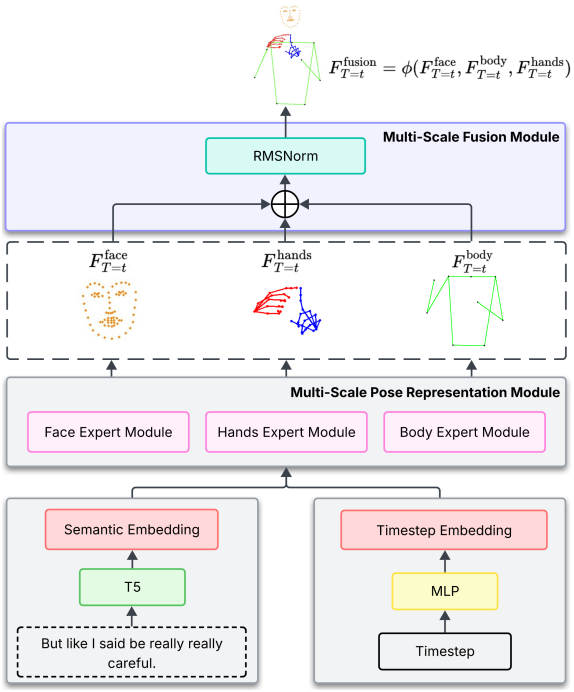


Figure 1: Overview of Hybrid Autoregressive-Diffusion Model.

## 4 Method

Our framework integrates a Hybrid Autoregressive-Diffusion model, Multi-Scale Pose Representation and Fusion modules, and a Confidence-Aware Causal Attention mechanism for temporally coherent, high-quality sign language pose generation.

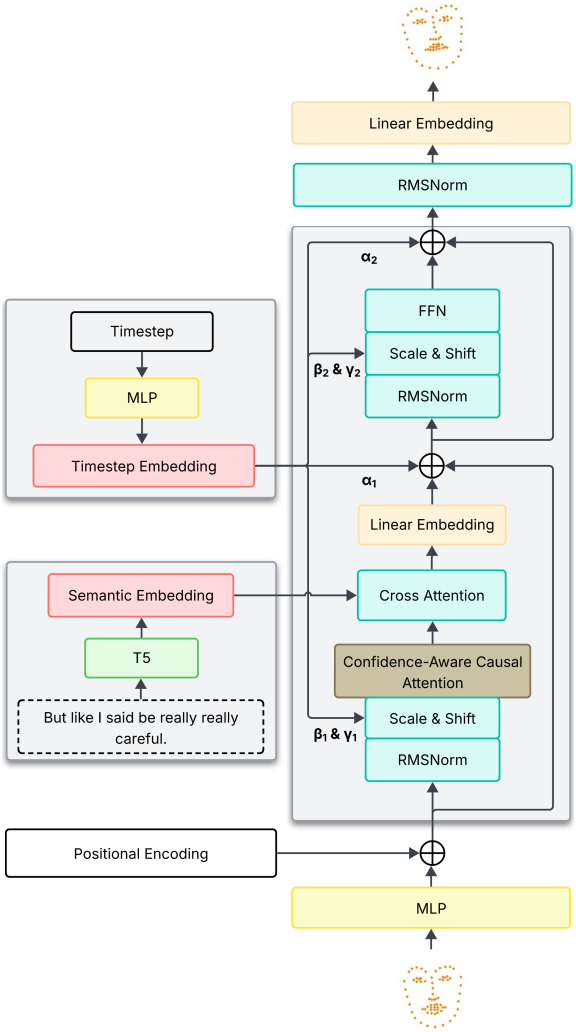


Figure 2: Expert module for facial features in the Multi-Scale Pose Representation Module.

### 4.1 Overview

As shown in Figure 1, given a natural language sentence, the model generates a temporally consistent pose sequence frame by frame. The Multi-Scale Pose Representation module first produces face, body, and hand articulators via dedicated Expert modules, which are fused into a full pose frame. The fused pose is then decomposed and used as the autoregressive condition for the next frame. Joint confidence scores guide the process to improve accuracy and realism.

### 4.2 Hybrid Autoregressive-Diffusion Model

To the best of our knowledge, this is the first SLP model that combines autoregressive and diffusion paradigms for SLP, aiming to harness the advantages of both.

Specifically, we introduce causal attention into the denoiser of the diffusion model by applying an attention mask $M \in \mathbb{R}^{n\times n}$ to its self-attention mechanism:

$$M_{ij} = \begin{cases} 0 & \text{if } j \le i \\ -\infty & \text{if } j > i \end{cases} \tag{10}$$

so that each position $i$ can only attend to positions $j \le i$ (i.e., current and previous tokens).

To address the distribution mismatch between training and inference commonly observed in autoregressive models, we draw on the training strategy proposed in Self-Forcing (Huang et al., 2025). Instead of relying on ground truth during training, our model always conditions the next step on its own previously generated articulators. Specifically, at each time step $t$, the model produces

$$\begin{aligned} \tilde{A}_t &= f_{\text{AR}}(\tilde{A}_{<t}), \\ \tilde{A}_{<t} &= \{\tilde{A}_1, \ldots, \tilde{A}_{t-1}\}, \\ \tilde{A}_t &= \{\tilde{F}_t, \tilde{H}_t, \tilde{B}_t\}. \end{aligned} \tag{11}$$

where $\tilde{F}_t$, $\tilde{H}_t$, and $\tilde{B}_t$ denote the generated face, hands, and body components at time step $t$, respectively. These components are fused into a full-frame pose and then decomposed again to serve as the condition for the next step. Using $\tilde{A}_t$ for the articulator tuple avoids overloading the confidence notation $C_t$ introduced in Section 3.3.

**Self-forcing protocol.** For each sentence, the first frame is generated from the text condition alone. For $t > 1$, the three experts generate the current face, hand, and body components in parallel, conditioned on the decomposed prediction from time $t-1$. During training, we do not replace this autoregressive condition with the ground truth. As a result, training and inference share the same conditioning path, which substantially reduces exposure bias. The holistic Soft-DTW loss (Cuturi and Blondel, 2018) is applied on the generated sequence to provide sequence-level supervision and further mitigate error accumulation.

To improve training efficiency and sampling speed, we adopt a flow-based diffusion model (Lipman et al., 2024, 2023) instead of the traditional DDPM-based approach (Ho et al., 2020; Song et al., 2020). Flow matching enables learning a continuous transformation path that maps the initial distribution directly to the target distribution, allowing sampling to be completed in fewer steps and enhancing both training and sampling efficiency.

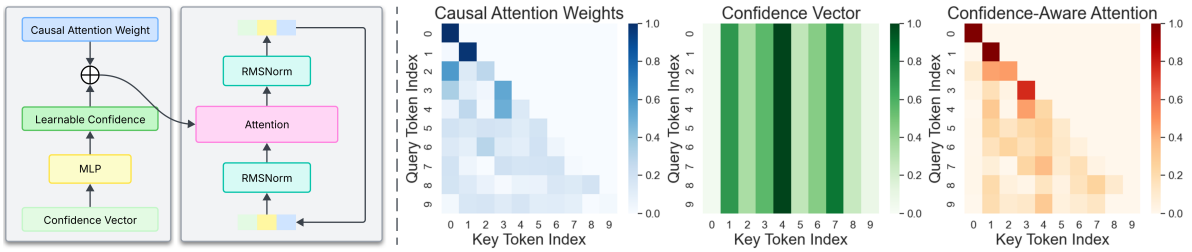


Figure 3: Confidence-Aware Causal Attention Mechanism.

### 4.3 Multi-Scale Pose Representation Module

Sign language involves coordinated motion across face, body, and hands. To capture this, we design a **Multi-Scale Pose Representation** module (Figure 2) that partitions keypoints into three anatomical groups and processes each with a dedicated expert module.

Given a sequence of $T$ frames with $N$ keypoints per frame, each keypoint $p_t^{(i)} = [x_t^{(i)}, y_t^{(i)}, c_t^{(i)}]$ is embedded as:

$$s_t^{(i)} = \text{MLP}(p_t^{(i)}) + PE(t), \tag{12}$$

We divide the keypoints into groups $\mathcal{G} = G_{\text{face}}, G_{\text{body}}, G_{\text{hands}}$ and extract group-specific features:

$$S_t^{(k)} = s_t^{(i)} \mid i \in G_k, \quad k = 1, 2, 3 \tag{13}$$

Each group sequence is processed by a scale-specific expert $\mathcal{E}^{(k)}$ to capture intra-scale dependencies:

$$H^{(k)} = \mathcal{E}^{(k)}([S_1^{(k)}, \ldots, S_T^{(k)}]) \tag{14}$$

We average-pool over joints and apply attention-based fusion across scales:

$$\begin{aligned} H_t &= \sum_{k=1}^{3} \alpha_t^{(k)} H_t^{(k)}, \\ \alpha_t^{(k)} &= \frac{\exp\left(w^\top \tanh\left(w_f H_t^{(k)}\right)\right)}{\sum_{j=1}^{3} \exp\left(w^\top \tanh\left(w_f H_t^{(j)}\right)\right)}. \end{aligned} \tag{15}$$

The fused representation $H_t$ serves as input for subsequent generation modules.

### 4.4 Confidence-Aware Causal Attention Mechanism

To improve robustness in autoregressive pose generation, we propose a **Confidence-Aware Causal Attention** module (Figure 3), which integrates keypoint confidence scores into the attention computation to downweight unreliable inputs.

Given a sentence $S$, the model generates a pose sequence $x = x^{(1)}, \ldots, x^{(T)}$, where each $x^{(t)} \in \mathbb{R}^{J \times d}$ represents $J$ keypoints. Each keypoint has an associated confidence score $c_i^{(t)} \in [0, 1]$.

**Confidence-Weighted Attention.** In causal attention, timestep $t$ only attends to $1, \ldots, t$:

$$\alpha_{t,s} = \frac{\exp\left(\frac{Q_t K_s^\top}{\sqrt{d_k}}\right)}{\sum_{j=1}^{t} \exp\left(\frac{Q_t K_j^\top}{\sqrt{d_k}}\right)} \tag{16}$$

We introduce a confidence bias to this formulation:

$$\alpha_{t,s} = \frac{\exp\left(\frac{Q_t K_s^\top}{\sqrt{d_k}} + \beta \cdot \bar{c}^{(s)}\right)}{\sum_{j=1}^{t} \exp\left(\frac{Q_t K_j^\top}{\sqrt{d_k}} + \beta \cdot \bar{c}^{(j)}\right)} \tag{17}$$

where $\bar{c}^{(s)} = \frac{1}{J} \sum_{i=1}^{J} c_i^{(s)}$ is the average keypoint confidence, and $\beta$ is a learnable scalar controlling the strength of the bias.

This allows the model to attend more to reliable frames during decoding.

### 4.5 Training Objectives

Our model is optimized using a composite loss that combines three complementary objectives: joint accuracy, kinematic consistency, and temporal alignment.

**Joint loss** To encourage accurate joint localization, we minimize the mean absolute error between predictions and ground truth:

$$\mathcal{L}_{\text{joint}} = \frac{1}{J} \sum_{j=1}^{J} \|p_j - \hat{p}_j\|_1, \tag{18}$$

| Methods | DEV | | | | | | TEST | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | B1↑ | B4↑ | ROUGE↑ | WER↓ | DTW↓ | FID↓ | B1↑ | B4↑ | ROUGE↑ | WER↓ | DTW↓ | FID↓ |
| PT (Saunders et al., 2020b) | 12.51 | 3.88 | 11.87 | 96.85 | - | - | 13.35 | 4.31 | 13.17 | 96.50 | - | - |
| G2P-DDM (Xie et al., 2024) | - | - | - | - | - | - | 16.11 | 7.50 | - | 77.26 | - | - |
| GCDM (Tang et al., 2025) | 22.88 | 7.64 | 23.35 | 82.81 | 11.18 | 39.87 | 22.03 | 7.91 | 23.20 | 81.94 | 11.10 | 49.22 |
| GEN-OBT (Tang et al., 2022) | 24.92 | 8.68 | 25.21 | 82.36 | - | - | 23.08 | 8.01 | 23.49 | 81.78 | - | - |
| Sign-IDD (Tang et al., 2024b) | 25.40 | 8.93 | 27.60 | 77.72 | 5.09 | 39.11 | 24.80 | 9.08 | 26.58 | 76.66 | 6.20 | 47.19 |
| **HybridSign (Ours)** | **26.98** | **9.26** | **28.07** | **75.81** | **4.07** | **38.28** | **25.77** | **10.03** | **27.97** | **75.02** | **4.96** | **45.50** |
| Ground Truth | 29.77 | 12.13 | 29.60 | 74.17 | 0.00 | 0.00 | 29.76 | 11.93 | 28.98 | 71.94 | 0.00 | 0.00 |

Table 1: Performance comparison on PHOENIX14T. Higher B1/B4/ROUGE indicates better back-translation quality, while lower WER/DTW/FID indicates better motion fidelity and temporal alignment.

| Methods | DEV | | | | | | TEST | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | B1↑ | B4↑ | ROUGE↑ | WER↓ | DTW↓ | FID↓ | B1↑ | B4↑ | ROUGE↑ | WER↓ | DTW↓ | FID↓ |
| PT (Saunders et al., 2020b) | 14.34 | 4.07 | 8.12 | 96.91 | 10.53 | 55.02 | 14.05 | 4.12 | 8.42 | 96.47 | 10.18 | 54.57 |
| G2P-DDM (Xie et al., 2024) | 19.82 | 5.37 | 12.47 | 90.05 | 8.25 | 50.48 | 19.48 | 5.12 | 12.21 | 89.58 | 7.97 | 49.83 |
| GCDM (Tang et al., 2025) | 26.43 | 5.84 | 15.53 | 91.92 | 6.32 | 46.19 | 25.91 | 5.57 | 15.21 | 91.43 | 6.13 | 45.71 |
| GEN-OBT (Tang et al., 2022) | 28.63 | 6.14 | 16.23 | 91.07 | 7.08 | 48.03 | 27.82 | 5.92 | 15.88 | 90.63 | 6.87 | 47.28 |
| Sign-IDD (Tang et al., 2024b) | 29.12 | 6.27 | 17.01 | 89.95 | 4.76 | 35.03 | 28.90 | 6.06 | 16.21 | 89.98 | 4.86 | 39.02 |
| **HybridSign (Ours)** | **30.71** | **6.92** | **18.96** | **87.11** | **3.76** | **34.19** | **30.12** | **6.48** | **18.02** | **88.30** | **3.89** | **37.10** |
| Ground Truth | 35.20 | 8.89 | 22.45 | 83.79 | 0.00 | 0.00 | 34.01 | 8.03 | 21.87 | 81.94 | 0.00 | 0.00 |

Table 2: Performance comparison on How2Sign. Higher B1/B4/ROUGE indicates better back-translation quality, while lower WER/DTW/FID indicates better motion fidelity and temporal alignment.

where $p_j \in \mathbb{R}^d$ and $\hat{p}_j \in \mathbb{R}^d$ denote the ground-truth and predicted position of joint $j$, and $\|\cdot\|_1$ is the $\ell_1$ norm.

**Bone loss** To preserve bone orientations and kinematic consistency, we penalize orientation deviations:

$$\mathcal{L}_{\text{bone}} = \frac{1}{B}\sum_{b=1}^{B} \|q_b - \hat{q}_b\|_2^2, \tag{19}$$

where $q_b$ and $\hat{q}_b$ are the ground-truth and predicted orientation (e.g. unit quaternions or axis-angle vectors) of bone $b$, and $\|\cdot\|_2$ denotes the Euclidean norm. If orientations are represented as quaternions, consider using a geodesic/angular metric as an alternative.

**Soft-DTW loss** To align predicted and ground-truth pose sequences temporally, we employ differentiable Dynamic Time Warping (Soft-DTW) (Cuturi and Blondel, 2018):

$$\mathcal{L}_{\text{soft-dtw}} = \min\nolimits_A^{\gamma} \sum_{(i,j)\in A} \|x_i - y_j\|_2^2, \tag{20}$$

where $x_i$ and $y_j$ are per-frame pose descriptors, $A$ is an alignment path, and $\gamma > 0$ controls the smoothness of the soft minimum:

$$\min\nolimits^{\gamma}\{a_1, \ldots, a_k\} = -\gamma \log\left(\sum_{i=1}^{k} e^{-a_i/\gamma}\right). \tag{21}$$

**Adaptive total loss** We balance the three terms with dynamic weights computed from the inverse exponential moving average (EMA) of each loss (Liu et al., 2019). Let $\bar{\mathcal{L}}_i^{(t)}$ be the EMA of loss $i$ at iteration $t$, and $\epsilon > 0$ a small constant to avoid division by zero. The weight for loss $i$ is

$$\lambda_i^{(t)} = \frac{\left(\bar{\mathcal{L}}_i^{(t)} + \epsilon\right)^{-1}}{\sum_{j=1}^{N}\left(\bar{\mathcal{L}}_j^{(t)} + \epsilon\right)^{-1}}, \qquad i = 1, \ldots, N, \tag{22}$$

with $N = 3$ in our case. The final objective at iteration $t$ is

$$\mathcal{L}_{\text{total}}^{(t)} = \lambda_1^{(t)}\,\mathcal{L}_{\text{joint}}^{(t)} + \lambda_2^{(t)}\,\mathcal{L}_{\text{bone}}^{(t)} + \lambda_3^{(t)}\,\mathcal{L}_{\text{soft-dtw}}^{(t)}. \tag{23}$$

This adaptive weighting scheme emphasizes losses that are relatively smaller (via inverse EMA), promoting stable and balanced training across spatial accuracy, kinematic consistency, and temporal alignment.

## 5 Experiment

### 5.1 Experimental Setup

**Datasets.** We evaluate the proposed method on two benchmarks: PHOENIX14T (Camgoz et al., 2018) and How2Sign (Duarte et al., 2021). PHOENIX14T is a German Sign Language corpus that consists of 8,257 sentence-level samples and 2,887 unique German words. How2Sign is a large-scale American Sign Language corpus designed for

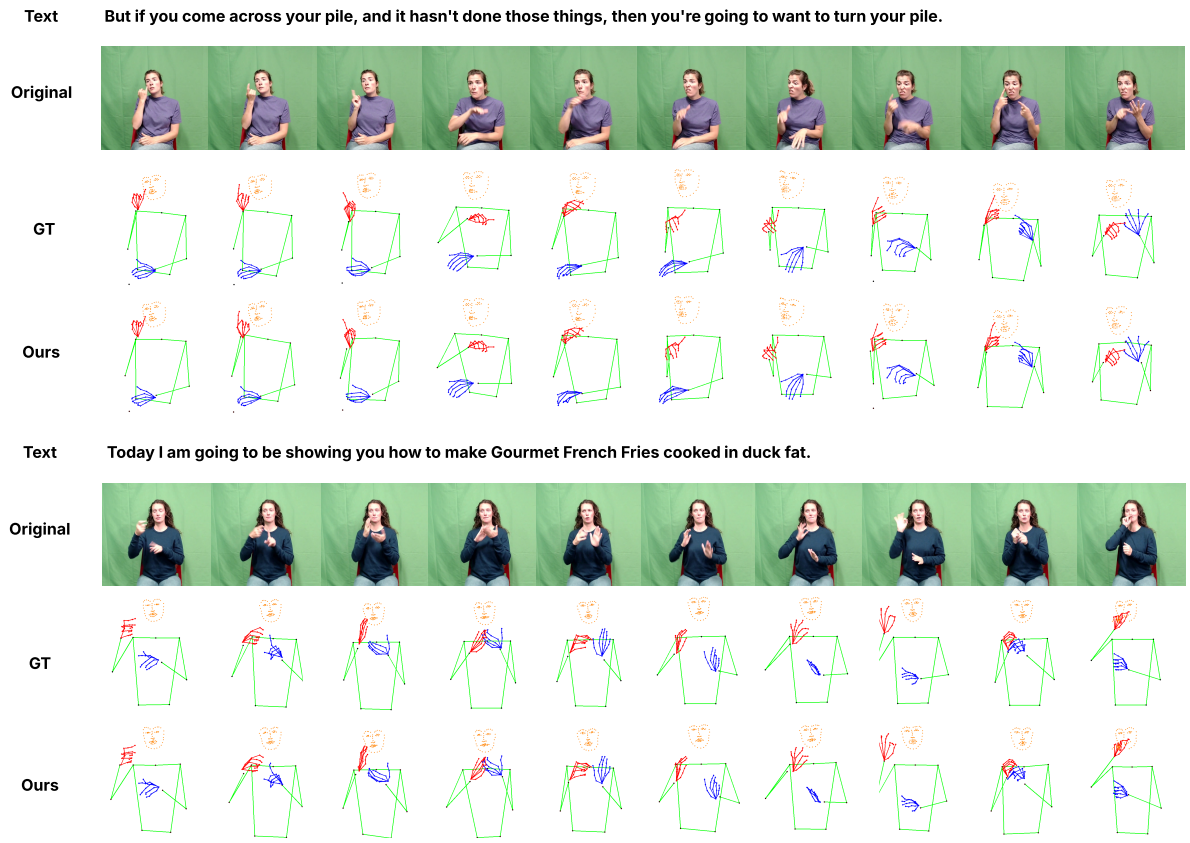


Figure 4: Visualization examples of generated poses on How2Sign. We compare HybridSign with the ground-truth poses and the original video frames for clear evaluation.

sign language understanding and generation tasks, containing more than 80 hours of multimodal data.

**Evaluation Metrics.** Following the existing works (Xie et al., 2024; Tang et al., 2025, 2024b,a), we used a pre-trained SLT model (Camgoz et al., 2020b) for back-translation, converting generated sign language pose sequences back to text, and then evaluated the results with BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), WER, DTW, and FID.

**Training and inference protocol.** Unless otherwise stated, all models generate sequences of 60 frames. The three articulator experts are executed in parallel within each time step, while the autoregressive dependency is imposed only across time. We report *time-to-first-frame* as latency, i.e., the wall-clock time required to output the first pose frame of a 60-frame sequence, because this quantity best reflects interactive usability. Throughout the paper, our use of the term *real-time* therefore refers to this low-latency, continuous generation setting relative to prior diffusion-based SLP systems rather than to instantaneous end-to-end video synthesis. During both training and inference, the previous pose fed to the next step is always the model prediction rather than the ground truth, which keeps the conditioning distribution consistent with the self-forcing design described in Section 4.2. Unless otherwise stated, the ablation results in Tables 4–6 are reported on the How2Sign test split.

## 5.2 Comparisons with Baseline Models

We conducted both qualitative and quantitative analyses to evaluate the model's performance.

| Method | Latency (s)↓ | Throughput (FPS)↑ |
|---|---|---|
| GCDM (Tang et al., 2025) | 52.18 | 1.15 |
| Sign-IDD (Tang et al., 2024b) | 40.31 | 1.49 |
| G2P-DDM (Xie et al., 2024) | 25.78 | 2.33 |
| **HybridSign** | **5.90** | **10.17** |

Table 3: Latency and throughput comparison under the 60-frame evaluation protocol. Latency measures time-to-first-frame rather than the total sequence generation time.

**Quantitative Analysis.** Tables 1 and 2 present a comparison of HybridSign and the baseline models on PHOENIX14T and How2Sign, respectively. Across both datasets, HybridSign delivers

the strongest overall quality–efficiency trade-off and achieves the best results among the compared methods on the reported metrics.

Table 3 compares latency and throughput with the baseline models. Here, latency refers to the time required to generate the first frame of the 60-frame sequence, rather than the total time to produce the entire video. This reflects the model's initial response speed, which is especially important for interactive deployment where a system should start responding before the entire sequence is synthesized. HybridSign achieves substantially lower latency and higher throughput than the compared diffusion-based baselines, indicating a strong quality–efficiency trade-off for low-latency sign language production.

Performance differences across the two datasets are relatively small, which is partly attributable to our use of a unified 2D pose representation across both benchmarks. This choice improves cross-dataset consistency and computational efficiency, but it also removes depth cues that are important for subtle hand-face interactions and for disambiguating overlapping articulators. Our Multi-Scale Pose Representation and Confidence-Aware Causal Attention mitigate this issue by emphasizing reliable local structure, yet they do not fully recover missing 3D information. Notably, the DTW score is reduced by approximately 20% thanks to the Soft-DTW loss (Cuturi and Blondel, 2018), which improves temporal alignment and stabilizes long-horizon autoregressive generation.

**Qualitative Analysis.** Figure 4 illustrates representative generated poses alongside the corresponding ground truth, and Figure 5 further expands the qualitative evaluation with more diverse examples. Across both figures, HybridSign preserves the global motion trend, hand trajectories, and major body posture changes. The most visible deviations occur in subtle bone lengths, fine wrist angles, and local configurations when multiple articulators become spatially close. Even in these cases, the generated poses remain semantically interpretable, which is consistent with the strong back-translation scores in Tables 1 and 2.

| Modes | B1↑ | B4↑ | DTW↓ | Latency (s)↓ | Throughput (FPS)↑ |
|---|---|---|---|---|---|
| Diffusion Mode | **30.25** | **6.55** | 8.06 | 32.89 | 1.83 |
| Autoregressive Mode | 26.15 | 5.40 | 4.49 | **5.53** | **10.85** |
| **Hybrid Mode** | 30.12 | 6.48 | **3.89** | 5.90 | 10.17 |

Table 4: Ablation study results of generation modes on the How2Sign test split.

| Modules | B1↑ | B4↑ | DTW↓ | Latency (s)↓ | Throughput (FPS)↑ |
|---|---|---|---|---|---|
| RNN | 23.47 | 5.02 | 6.89 | **4.72** | 9.71 |
| CA† | 25.08 | 5.83 | 5.50 | 5.48 | **10.95** |
| **CACA‡** | **30.12** | **6.48** | **3.89** | 5.90 | 10.17 |

Table 5: Ablation study results of autoregressive backbones on the How2Sign test split. CA† denotes Causal Attention, and CACA‡ denotes Confidence-Aware Causal Attention.

| Experts | B1↑ | B4↑ | DTW↓ | Latency (s)↓ | Throughput (FPS)↑ |
|---|---|---|---|---|---|
| 1 (whole pose) | 22.33 | 5.14 | 7.02 | 7.69 | 7.80 |
| 4 (face + body + lh + rh) | 29.17 | 6.03 | 5.72 | 7.73 | 7.76 |
| **3 (face + body + hands)** | **30.12** | **6.48** | **3.89** | **5.90** | **10.17** |

Table 6: Ablation study results of different expert decompositions on the How2Sign test split.

### 5.3 Ablation Study

We conducted three ablation studies on the generation of 60-frame sign language pose sequences to systematically evaluate our method. Unless otherwise stated, the ablations are reported on the How2Sign test split. Specifically, we compared the proposed hybrid approach with a pure diffusion model and a pure autoregressive model, and further investigated the impact of different attention mechanisms and varying numbers of expert modules.

Table 4 compares the performance and efficiency of a pure diffusion model and a pure autoregressive model. Specifically, the diffusion model adopts the DiT architecture (Peebles and Xie, 2023), while the autoregressive model uses a Transformer decoder with teacher forcing (Vaswani et al., 2017). Experimental results show that, thanks to its multi-step denoising process, the diffusion model achieves very high generation quality and reaches the best scores on both the B1 and B4 metrics. However, its temporal consistency and generation efficiency lag significantly behind the other models, making it unsuitable for low-latency scenarios. In contrast, the autoregressive model generates frame-by-frame, resulting in extremely low first-frame latency, but suffers from a noticeable quality drop due to the mismatch between training and inference. Our proposed approach combines the high-quality generation of diffusion models with the frame-by-frame generation of autoregressive models, achieving both strong performance and efficiency.

Table 5 presents the performance differences among various autoregressive implementations, including a recurrent neural network

(RNN) (Graves, 2012), a standard causal attention mechanism (Vaswani et al., 2017), and a confidence-aware causal attention mechanism. The results indicate that the RNN-based model achieves very low generation quality, but due to its computational simplicity compared to attention mechanisms, it exhibits excellent first-frame latency. However, its strict sequential dependency limits overall efficiency, resulting in lower throughput than attention-based methods. In contrast, the confidence-aware causal attention mechanism incorporates confidence scores, assigning different attention weights based on the reliability of keypoints, which leads to the best generation quality while maintaining strong applicability in low-latency scenarios.

Table 6 presents the results of varying the number of expert modules. The experiments include using a single expert module (treating the entire pose as a whole), using three expert modules to handle the face, body, and hands separately, and using four expert modules to handle the face, body, left hand, and right hand individually. The single-expert setting cannot capture local articulator-specific details, leading to a significant drop in generation quality. More importantly, the comparison between three and four experts reveals that finer decomposition is not always better. In sign language, the two hands form a strongly coupled subsystem: many signs rely on symmetry, anti-symmetry, relative hand distance, and precise cross-hand timing. A unified hands expert can model these bimanual dependencies directly within one latent space. By contrast, splitting the hands into two experts forces the fusion stage to reconstruct cross-hand relations only after separate generation, which weakens explicit modeling of relative geometry and increases the chance of temporal misalignment. This explains why the four-expert variant underperforms the three-expert design in both quality and efficiency. Therefore, three experts provide the best trade-off, balancing local specialization with coherent bimanual coordination.

### 5.4 Discussion on 2D Pose Representation

Our model adopts 2D poses as a pragmatic representation to keep preprocessing consistent across PHOENIX14T and How2Sign and to avoid introducing dataset-specific 3D supervision requirements. This design is effective for large-scale benchmarking, but it also exposes a clear limitation: depth ambiguity is collapsed in the 2D projection. In practice, the most challenging cases arise when the hands approach the face, when one hand occludes the other, or when similar 2D projections correspond to different 3D articulations. The qualitative examples in Figures 4 and 5 suggest that HybridSign is robust for dominant planar motion, but still exhibits larger local deviations in precisely these ambiguous cases. We therefore view 3D-aware sign production as an important next step: our hybrid autoregressive-diffusion framework is agnostic to the pose dimensionality and can naturally benefit from richer 3D or multi-view pose annotations when such data become available.

## 6 Conclusions

We introduce a hybrid autoregressive-diffusion framework for low-latency SLP that combines temporal modeling with high-quality refinement. Multi-scale pose representation and confidence-aware causal attention improve both accuracy and robustness. Experiments on PHOENIX14T and How2Sign validate the effectiveness of the approach in both generation quality and efficiency under a low-latency evaluation setting.

## 7 Limitations

While our hybrid autoregressive-diffusion framework achieves state-of-the-art performance for low-latency SLP, several limitations remain. First, the approach depends on annotated sign language datasets, which are still limited in size and diversity, potentially restricting generalization to less-represented sign languages or signing styles. Second, although the multi-scale representation captures hand and facial details, subtle finger articulations and non-manual signals (e.g., eye gaze and mouth gestures) are not yet fully modeled. Finally, while the current inference speed is promising for interactive applications, further optimization may be required for deployment on resource-constrained devices.

## Acknowledgments

Manoharan and Ye were supported in part by the Smart Ideas (UOA2493, Developing a Reo Turi Interpreter for Ngati Turi/Sign Language Interpreter Using Weighted Multimodal Network for Mahuta ki Tai) funded by the MBIE, New Zealand.

## References


Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020a. Multi-channel transformers for multi-articulatory sign language translation. *Preprint*, arXiv:2009.00299.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020b. Sign language transformers: Joint end-to-end sign language recognition and translation. *Preprint*, arXiv:2003.13830.

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Marco Cuturi and Mathieu Blondel. 2018. Soft-dtw: a differentiable loss function for time-series. *Preprint*, arXiv:1703.01541.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sen Fang, Chunyu Sui, Yanghao Zhou, Xuedong Zhang, Hongbin Zhong, Yapeng Tian, and Chen Chen. 2025. Signdiff: Diffusion model for american sign language production. *Preprint*, arXiv:2308.16082.

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. Llms are good sign language translators. *Preprint*, arXiv:2404.00925.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *Preprint*, arXiv:1211.3711.

Dan Guo, Shengeng Tang, and Meng Wang. 2019. Connectionist temporal modeling of video and language: a joint model for translation and sign labeling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 751–757.

Dan Guo, Wengang Zhou, Houqiang Li, and Meng Wang. 2017. Online early-late fusion based on adaptive hmm for sign language recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(1):1–18.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Preprint*, arXiv:2006.11239.

Hezhen Hu, Junfu Pu, Wengang Zhou, and Houqiang Li. 2023a. Collaborative multilingual continuous sign language recognition: A unified framework. *IEEE Transactions on Multimedia*, 25:7559–7570.

Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. 2023b. Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11221–11239.

Wencan Huang, Wenwen Pan, Zhou Zhao, and Qi Tian. 2021. Towards fast and high-quality sign language production. *Proceedings of the 29th ACM International Conference on Multimedia*.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. 2025. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*.

Oscar Koller. 2020. Quantitative survey of the state of the art in sign language recognition. *Preprint*, arXiv:2008.09918.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow matching for generative modeling. *Preprint*, arXiv:2210.02747.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. 2024. Flow matching guide and code. *Preprint*, arXiv:2412.06264.

Shikun Liu, Edward Johns, and Andrew J. Davison. 2019. End-to-end multi-task learning with attention. *Preprint*, arXiv:1803.10704.

Xingchao Liu, Chengyue Gong, and Qiang Liu. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. *Preprint*, arXiv:2209.03003.

Seshadri Mazumder, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. 2021. Translating sign language videos to talking faces. In *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*.

John McDonald, Rosalee J. Wolfe, Jerry Schnepp, Julie Hochgesang, Diana Gorman Jamrozik, Marie Stumbo, Larwan Berke, Melissa Bialek, and Farah Thomas. 2016. An automated technique for real-time production of lifelike animations of american sign language. *Universal Access in the Information Society*, 15(4):551–566.

Alex Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. *Preprint*, arXiv:2102.09672.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. *Preprint*, arXiv:2212.09748.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020a. Adversarial training for multi-channel sign language production. *Preprint*, arXiv:2008.12405.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020b. Progressive transformers for end-to-end sign language production. *Preprint*, arXiv:2004.14874.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks. *Preprint*, arXiv:2103.06982.

Jérémie Segouat. 2009. A study of sign language coarticulation. *ACM Sigaccess Accessibility and Computing*, 93.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *Preprint*, arXiv:2010.02502.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-based generative modeling through stochastic differential equations. *Preprint*, arXiv:2011.13456.

Stephanie Stoll, Necati Camgoz, Simon Hadfield, and Richard Bowden. 2020. Text2sign: Towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*.

Shengeng Tang, Jiayi He, Lechao Cheng, Jingjing Wu, Dan Guo, and Richang Hong. 2024a. Discrete to continuous: Generating smooth transition poses from sign language observation. *Preprint*, arXiv:2411.16810.

Shengeng Tang, Jiayi He, Dan Guo, Yanyan Wei, Feng Li, and Richang Hong. 2024b. Sign-idd: Iconicity disentangled diffusion for sign language production. *Preprint*, arXiv:2412.13609.

Shengeng Tang, Richang Hong, Dan Guo, and Meng Wang. 2022. Gloss semantic-enhanced network with online back-translation for sign language production. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 5630–5638, New York, NY, USA. Association for Computing Machinery.

Shengeng Tang, Feng Xue, Jingjing Wu, Shuo Wang, and Richang Hong. 2025. Gloss-driven conditional diffusion models for sign language production. *ACM Trans. Multimedia Comput. Commun. Appl.*, 21(4).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Preprint*, arXiv:1706.03762.

Pan Xie, Qipeng Zhang, Peng Taiying, Hao Tang, Yao Du, and Zexian Li. 2024. G2p-ddm: Generating sign pose sequence from gloss sequence with discrete diffusion model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6).

Weichao Zhao, Hezhen Hu, Wengang Zhou, Jiaxin Shi, and Houqiang Li. 2023. Best: Bert pre-training for sign language recognition with coupling tokenization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3597–3605.

## A Additional Implementation and Evaluation Details

This appendix provides supplementary details on data preprocessing, training configuration, evaluation protocols, and additional qualitative results.

## B Datasets and Preprocessing

**Datasets.** We use two benchmark datasets: PHOENIX14T (Camgoz et al., 2018) and How2Sign (Duarte et al., 2021). We follow the official train/validation/test splits for both datasets.

| Dataset | Language | Gloss annotations | Continuous signing | Native pose keypoints |
|---|---|---|---|---|
| PHOENIX14T | German | Yes | Yes | No |
| How2Sign | English | Yes | Yes | Yes |

Table 7: Overview of the datasets used in this work. "Native pose keypoints" indicates whether pose annotations are provided by the original dataset release.

**Pose extraction and representation.** PHOENIX14T does not provide pose sequences, whereas How2Sign includes pre-extracted 2D pose data. To keep the representation consistent across datasets, we extract 2D keypoints from PHOENIX14T video frames using OpenPose (Cao et al., 2019). The final pose representation contains 137 keypoints per frame, covering the body, hands, and face. Each keypoint is represented by its 2D coordinates together with a confidence score.

**Normalization.** To improve training stability, we normalize the $x$ and $y$ coordinates of all keypoints separately. Confidence scores are already bounded in $[0, 1]$ and are therefore kept unchanged. These

confidence values are later used by the Confidence-Aware Causal Attention module described in Section 4.4.

## C Training Details and Hyperparameters

**Hardware and runtime.** Training is performed separately on the two datasets using a single 80GB NVIDIA A100 GPU. On PHOENIX14T, the model converges in approximately 40 hours; on How2Sign, training requires approximately 45 hours under the same hardware setup.

**Optimization setup.** Unless otherwise stated in the main text, all experiments use the hyperparameters listed in Table 8. The default configuration uses three experts and bfloat16 precision, matching the setting reported for the main results.

## D Evaluation Metrics

Standard text-generation metrics cannot be applied directly to sign pose sequences. Following prior work, we therefore use a pre-trained sign language translation model for back-translation, convert the generated pose sequence back into text, and report both text-based and motion-based metrics. For efficiency reporting, latency denotes time-to-first-frame and throughput denotes the average generated frames per second under the same 60-frame protocol used in the main paper.

### D.1 Text-based Metrics

**BLEU.** BLEU (Papineni et al., 2002) measures $n$-gram precision between the back-translated sentence and the reference sentence. We report BLEU-1 and BLEU-4 to capture unigram and 4-gram overlap, respectively. Higher BLEU indicates better semantic agreement with the source sentence.

**ROUGE-L.** ROUGE (Lin, 2004) is a recall-oriented metric. We use ROUGE-L, which is based on the longest common subsequence between the back-translated sentence and the reference. Higher ROUGE-L indicates stronger overlap in sentence content and ordering.

**WER.** Word Error Rate (WER) measures the discrepancy between the back-translated sentence and the original reference:

$$\mathrm{WER} = \frac{S + D + I}{N}, \tag{24}$$

where $S$, $D$, and $I$ denote the numbers of substitutions, deletions, and insertions, and $N$ is the number of reference words. Lower WER indicates better preservation of the input semantics.

### D.2 Motion-based Metrics

**DTW.** Dynamic Time Warping (DTW) measures the alignment cost between two motion sequences that may differ in temporal dynamics:

$$\mathrm{DTW}(X, Y) = \min_{\pi} \sum_{(i,j)\in\pi} \|x_i - y_j\|_2, \tag{25}$$

where $\pi$ denotes a valid warping path. In our setting, DTW is computed over pose trajectories to evaluate structural and temporal similarity between generated and ground-truth motion. Lower DTW indicates better temporal alignment.

**FID.** Fréchet Inception Distance (FID) evaluates the distributional similarity between real and generated samples in a feature space:

$$\mathrm{FID} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right), \tag{26}$$

where $(\mu_r, \Sigma_r)$ and $(\mu_g, \Sigma_g)$ are the means and covariance matrices of real and generated samples, respectively. In our experiments, FID is adapted to pose-based sign generation by extracting features from keypoint sequences. Lower FID indicates more realistic generation.

## E Additional Qualitative Examples

Figure 5 provides additional qualitative examples on How2Sign. Compared with Figure 4 in the main paper, these examples cover more diverse motion patterns and lexical content, further illustrating that HybridSign preserves temporally coherent hand–body coordination while staying close to the reference pose sequence.

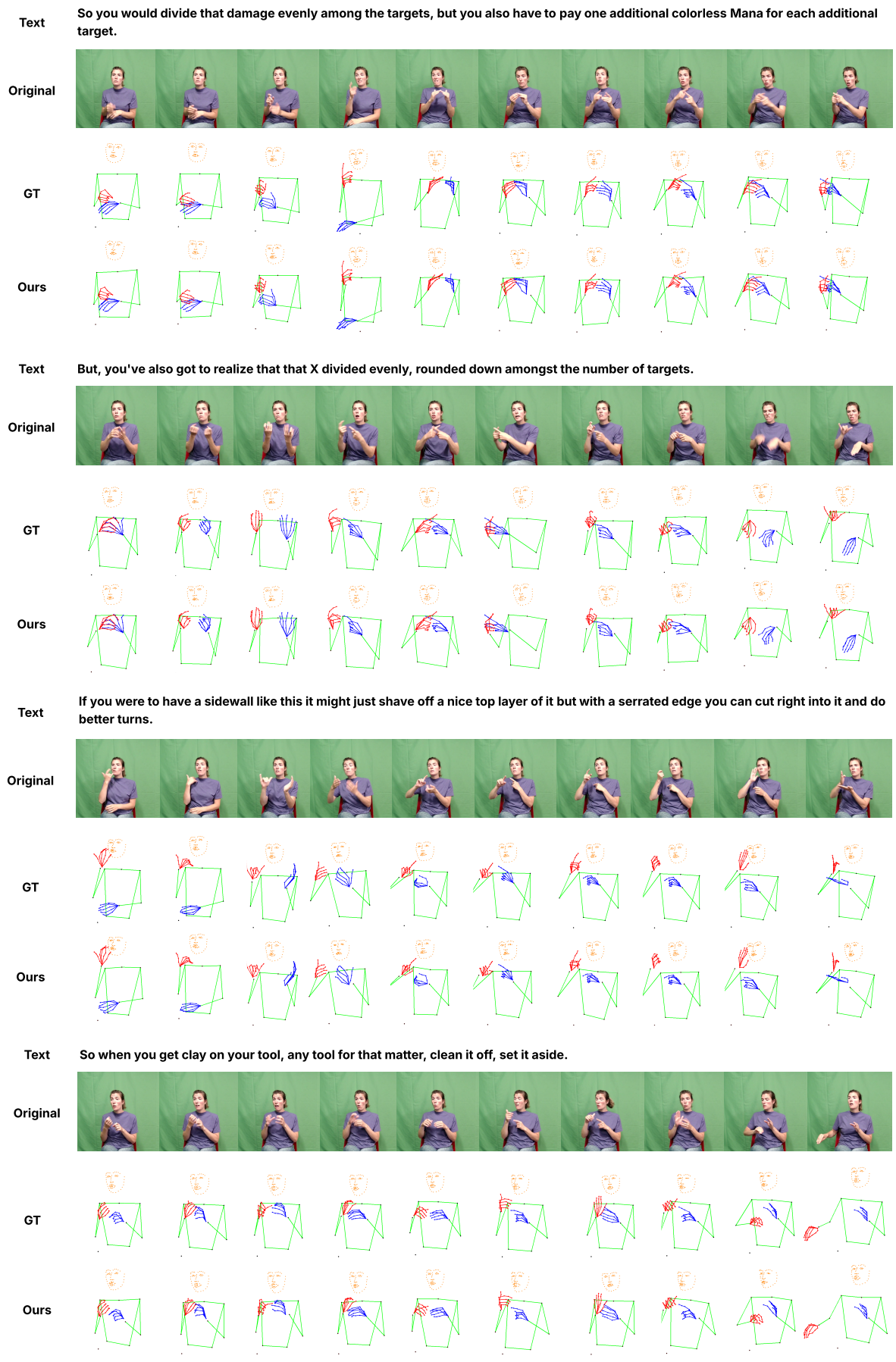


Figure 5: Additional qualitative examples on How2Sign. Compared with Figure 4, these samples cover more diverse motion patterns and lexical contents, showing that HybridSign remains close to the reference sequence while preserving temporally coherent hand–body coordination.

| Hyperparameter | PHOENIX14T | How2Sign |
|---|---|---|
| Seed | 42 | 42 |
| Optimizer | AdamW | AdamW |
| Learning rate | 1e-4 | 1e-4 |
| Weight decay | 0.01 | 0.01 |
| Scheduler | ExponentialLR | ExponentialLR |
| Decay factor | 0.5 | 0.1 |
| Batch size | 8 | 16 |
| Epochs | 5 | 3 |
| Dropout | 0.1 | 0.2 |
| Embedding dimension | 512 | 768 |
| Numerical precision | bfloat16 | bfloat16 |
| Number of experts | 3 | 3 |

Table 8: Training hyperparameters used in our experiments.