Stable Signer: Hierarchical Sign Language Generative Model

Sen Fang[1], **Yalin Feng**[2*], **Hongbin Zhong**[3], **Yanxin Zhang**[4], **Dimitris N. Metaxas**[1]
[1]Rutgers University, [2]Nanyang Technological University, [3]Georgia Institute of Technology,
[4]University of Wisconsin-Madison
https://stablesigner.github.io/

Abstract

Sign Language Production (SLP) is the process of converting the complex input text into a real video. Most previous works focused on the Text2Gloss, Gloss2Pose, Pose2Vid stages[1], and some concentrated on Prompt2Gloss and Text2Avatar stages. However, this field has made slow progress due to the inaccuracy of text conversion, pose generation, and the rendering of poses into real human videos in these stages, resulting in gradually accumulating errors. Therefore, in this paper, we streamline the traditional redundant structure, simplify and optimize the task objective, and design a new sign language generative model called **Stable Signer**. It redefines the SLP task as a hierarchical generation end-to-end task that only includes text understanding (Prompt2Gloss, Text2Gloss) and Pose2Vid, and executes text understanding through our proposed new **S**ign **L**anguage **U**nderstanding **L**inker called **SLUL**, and generates hand gestures through the named **SLP-MoE** hand gesture rendering expert block to end-to-end generate high-quality and multi-style sign language videos. SLUL is trained using the newly developed **S**emantic-**A**ware **G**loss **M**asking Loss (**SAGM Loss**). Its performance has improved by 48.6% compared to the current SOTA generation methods, which is a significant increase in the SLP field.

1 Introduction

In the field of sign language based on deep learning, Sign Language Recognition (SLR (Hu et al., 2023; Jiang et al., 2021; Tarrés et al., 2023)) first developed and became popular, and then around 2015-2020, Sign Language Production (SLP (Saunders et al., 2021c,a; Fang et al., 2025a)) also gradually became popular. However, compared to the mature and popular SLR and understanding tasks, *sign language production is relatively lagging behind*. This is caused by several reasons: **(1)** The processing methods for different datasets vary, and there are differences in accuracy, which makes the learning of postures relatively difficult. **(2)** Different generation or evaluation models are specific to certain posture formats, making the comparison and transfer relatively difficult. **(3)** The multi-stage SLP pipeline suffers from *error accumulation across stages*, making performance improvements challenging.

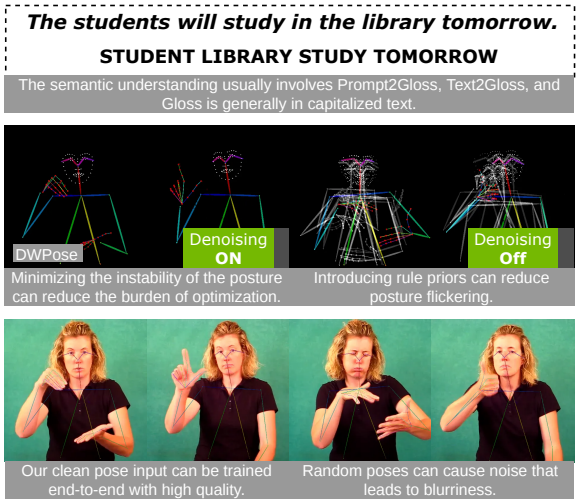


Figure 1: **The current SLP paths and their drawbacks:** SLP contains that the prompt/complex text gradually transforms into a pose video, and then it goes through the process of being converted into a real person video. However, this process is too complex to involves a lot of errors, which accumulate. Therefore, we plan to reduce the redundancy of unnecessary intermediate steps (as shown in the figure, same "Gloss" has poses that do not correspond in time), make the initial and final steps more closely connected, and then achieve end-to-end hierarchical model learning.

To address these issues, drawing on some of the similar work (Shazeer et al., 2017; Riquelme et al., 2021) in Sign Language Production (SLP), we realized that there is *redundancy in the process of SLP based on deep learning*: the input and output of the SLP task are text and the final high-quality video. For pose video, **rules can be used as a**

*Equal Contribution.
[1]Gloss represents the specific gesture text words.

prior (Stoll et al., 2020a), because the gloss and the text only need to correspond to a specific gesture or movement. The SLR cannot introduce rule priors because there are multiple possible inputs of similar gestures, while in SLP, as long as the precise gesture output is required, *the core problem of SLP is actually that the input of the gloss needs to be accurate*, so the pressure lies in the language understanding (Saharia et al., 2022) part.

And what's more serious is that the SLP methods actually did not take into account that their requirements are different from those of SLR and SLT. As shown in Fig 1, for a specific posture, there are *multiple expressions*. If we forcibly learn the generation of the posture, we are likely to fall into the *average posture (suboptimal state)*. If our training goals or data are not always precisely aligned, the more postures there are, the less accurate or difficult it is to learn at same time. Therefore, the motivation of end-to-end and the two existing problems prompt us to **streamline this redundant part** (Saunders et al., 2020b, 2021b). Then we can focus on researching and improving the truly necessary parts of language understanding (Saharia et al., 2022) and Pose2Vid.

Since we can now learn SLP in an rarely end-to-end manner, we need to consider the challenges that text understanding and pose-conditioned video generation will face: **(1)** Traditional SLP uses Gloss or Text as input (Zelinka and Kanis, 2020). To ensure accuracy, they also use translation models to translate Prompts or complex Text into Gloss, but it is not as thorough as ours. **(2)** When dealing with processing, many previous methods, in order to handle heterogeneous sign language data, repeatedly train the model or specially create methods, which unnecessarily increases the complexity of the task. **(3)** For the Pose2Video (Chan et al., 2019a; Saunders et al., 2020a) part, previous methods, due to the lack of end-to-end training, have difficulty using complex textual information for assistance or combining the pose input of Pose2Vid model more closely. This is the main reason why it is difficult to improve the final video effect.

To address these challenges, we uniformly set the core as a *rule-based learnable hybrid* Gloss2Pose[2]. Then, we designed a sign language understanding linker called **SLUL**, which can uniformly translate prompts or texts into Gloss. It is trained using the Prompt Mask Loss developed by us. This new Loss is used to handle different language inputs and prevent Gloss conflicts. Subsequently, we developed the **SLP-MoE module**, which inputs the semantic information of different language inputs into the Pose2Vid model's input for end-to-end training. Our method can utilize the semantic information and more posture information that were not utilized by previous methods, resulting in *higher video quality and better semantic expression*.

In summary, our contributions can be summarized as follows:

- A **Sign Language Understanding Linker (SLUL)** for unified understanding, which can accurately convert complex prompts or textual sign language information into Glosses required for different sign languages.
- **Sign Language Production Mixture-of-Experts (SLP-MoEs)**, which can receive output from SLUL and rapidly and accurately generate precise pose videos from Glosses, subsequently rendering them into diverse high-quality sign language videos.
- A **Semantic-Aware Gloss Masking (SAGM) Loss**, which enables end-to-end training on 2D sign language videos. Experiments demonstrate a *48.6% improvement* over the state-of-the-art, which represents a substantial advancement in the SLP domain.

2 Methodology

2.1 Data Construction

For the sign language, we use a dataset called Prompt2Sign (Fang et al., 2025a). The videos in this dataset are sourced from existing mainstream datasets and online sign language videos, and they are segmented into tens of thousands of clips. It has LLM-generated prompts, texts, and comprehensive glosses. It also provides a unified processing procedure to facilitate our comparison with existing methods. **For pose condition control**, we first use a video dataset called OpenVidHD[3] (Nan et al., 2025) for training to improve video quality. Then, we use sign language videos to enhance the posture guidance ability.

[2]We introduce prior knowledge not by completely giving up learning, but by ingeniously transforming the learning objective into achieving the optimal posture and optimizing the smoothness of transitions.

[3]https://github.com/NJU-PCALab/OpenVid-1M

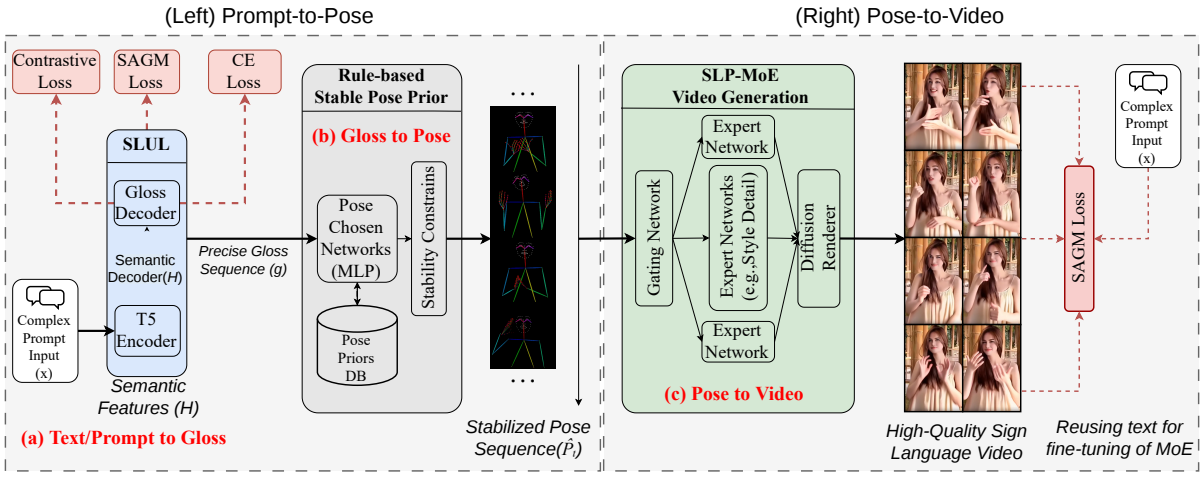


Figure 2: **Overview of the Prompt2Pose and Pose2Video pipeline: (a) Text/Prompt to Gloss:** The SLUL module uses a T5 encoder to process complex prompts and generate precise gloss sequences, trained with SLUL Loss, SAGM Loss, KL divergence, and contrastive loss. **(b) Gloss to Pose:** A rule-based stable pose prior database provides candidate poses, which are selected by the Pose Chosen Networks (MLP) guided by semantic features. The Gating Network and Expert Networks in SLP-MoE determine the optimal pose selection with stability constraints applied. **(c) Pose to Video:** The stabilized pose sequence is fed into a diffusion renderer to generate high-quality sign language videos in multiple styles, with the semantic features reused for fine-tuning the MoE module.

2.2 Linking Semantics with Sign Language Understanding

We cast prompt/text $\mathbf{x}$ (tagged by language identifier ℓ) into a sequence-to-sequence task that predicts a gloss sequence $\mathbf{g} = (g_1, \ldots, g_T)$. A T5 encoder produces contextual states $H = f_{\theta_E}([\ell; \mathbf{x}]) \in \mathbb{R}^{L \times d}$, where L is the input sequence length and d is the hidden dimension; the decoder autoregressively generates gloss tokens with maximum-likelihood training

$$\mathcal{L}_{\text{SLUL}} = -\sum_{t=1}^{T} \log p_\theta(g_t \mid g_{<t}, H).$$

This base term aligns prompts with gloss semantics while sharing encoder parameters.

Semantic Aware Gloss Masking Loss (SAGM Loss). To reduce gloss ambiguity and force semantic reconstruction, we randomly mask gloss tokens (mask rate ρ), forming $\tilde{\mathbf{g}}$ and requiring the model to infer the masked entries. The masked objective is

$$\mathcal{L}_{\text{SAGM}} = -\sum_{t} \mathbf{1}[u_t \le \rho] \, \log p_\theta(g_t \mid \tilde{\mathbf{g}}, H),$$

where $u_t \sim \mathcal{U}(0,1)$. We further enforce consistency between masked/unmasked posteriors

$$\mathcal{L}_{\text{KL}} = \text{KL}\big(p_\theta(\cdot \mid \tilde{\mathbf{g}}, H) \,\|\, p_\theta(\cdot \mid \mathbf{g}, H)\big),$$

so the decoder cannot drift when surface cues vanish. Intuitively, SAGM behaves as a semantic denoiser that combats noisy or rare gloss forms.

Stable SLUL and Contrastive Linking. We prepend ℓ to inputs and couple prompt/gloss embeddings via contrastive alignment:

$$\mathcal{L}_{\text{con}} = -\log \frac{\exp(\langle \bar{h}_x, \bar{h}_g \rangle / \tau)}{\sum_{g'} \exp(\langle \bar{h}_x, \bar{h}_{g'} \rangle / \tau)},$$

with mean-pooled embeddings $\bar{h}_x = \text{Pool}(H)$, $\bar{h}_g = \text{Pool}(f_{\theta_D}(\mathbf{g}))$ from encoder and decoder outputs respectively, and temperature τ. Following standard practice, negative samples g' are drawn from the same batch. The semantic objective becomes

$$\mathcal{L}_{\text{SLUL+SAGM}} = \mathcal{L}_{\text{SLUL}} + \lambda_{\text{SAGM}} \mathcal{L}_{\text{SAGM}} + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}} + \lambda_{\text{con}} \mathcal{L}_{\text{con}}.$$

This stack enforces (i) faithful prompt-to-gloss mapping, (ii) robustness to masked cues, and (iii) cross-lingual embedding agreement.

Pseudo-code (semantic stage). This semantic stage equips SLUL with robustness to noisy or under-specified prompts: masking forces the decoder to rely on encoder semantics, KL keeps masked/unmasked posteriors aligned, and contrastive linking ties prompt and gloss embeddings across languages. However, accurate gloss alone does not guarantee stable video; pose retrieval and temporal stability are critical. We therefore turn to a gated mixture-of-experts to select and blend pose priors before rendering.

Algorithm 1 SLUL + SAGM forward

$H \leftarrow f_{\theta_E}([\ell; \mathbf{x}])$
Sample masks $m_t \sim \text{Bernoulli}(\rho)$
Set $\tilde{g}_t \leftarrow$ `[MASK]` if $m_t = 1$ else g_t
Compute $p_\theta(\cdot \mid \tilde{\mathbf{g}}, H)$ and $p_\theta(\cdot \mid \mathbf{g}, H)$
$\mathcal{L}_{\text{SLUL+SAGM}} \leftarrow \mathcal{L}_{\text{SLUL}} + \lambda_{\text{SAGM}}\mathcal{L}_{\text{SAGM}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}$

2.3 SLP-MoE for Stable Pose2Video Production

Given predicted gloss $\mathbf{g}$ and semantic features H from SLUL, we select pose priors from a rule-based database with a gated mixture-of-experts (MoE), then stabilize keypoints before rendering.

SLP-MoE: Gated Pose Experts. A gloss-conditioned query $q = \text{Pool}(H)$ produces gates over K experts:

$$w_k = \frac{\exp(q^\top W_k)}{\sum_{j=1}^{K} \exp(q^\top W_j)},$$
$$\mathbf{p}_{\text{pose}} = \sum_{k=1}^{K} w_k \, \phi_k(\mathbf{g}),$$

where each expert ϕ_k retrieves a candidate pose sequence from the database based on gloss $\mathbf{g}$, and $\mathbf{p}_{\text{pose}}$ represents the weighted blend of these retrieved sequences. During training with ground truth pose sequence index y, we optimize

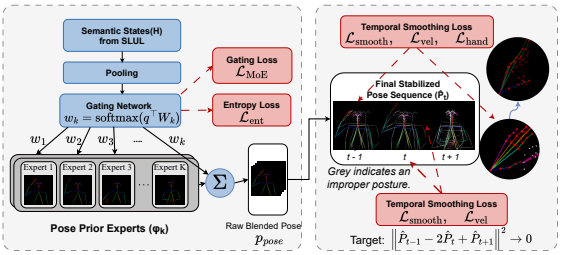
$$\mathcal{L}_{\text{MoE}} = -\log \sum_k w_k \mathbf{1}[k = y],$$
$$\mathcal{L}_{\text{ent}} = -\sum_k w_k \log w_k,$$

encouraging correct selection and avoiding gate collapse. At inference, we use the expert with highest gate weight.

Stability-Constrained Pose Blending. We refine the blended pose sequence $\mathbf{p}_{\text{pose}}$ into stabilized keypoints $\hat{P}_t \in \mathbb{R}^{J\times 2}$ for J body keypoints at frame t, extracting hand keypoint subset $\hat{H}_t$. We penalize jerk and preserve hand fidelity:

$$\mathcal{L}_{\text{smooth}} = \sum_t \|\hat{P}_t - 2\hat{P}_{t-1} + \hat{P}_{t-2}\|_2^2,$$
$$\mathcal{L}_{\text{hand}} = \sum_t \|\hat{H}_t - H_t^\star\|_2^2,$$

where $H_t^\star$ denotes ground truth hand keypoints. We optionally damp micro-flicker via $\mathcal{L}_{\text{vel}} = \sum_t \|\hat{P}_t - \hat{P}_{t-1}\|_2^2$. These terms directly target temporal jitter and hand intelligibility.



(a) MoE Gating & Blending Mechanism (b) Spatio-Temporal Stabilization Constraints

Figure 3: **Details of the SLP-MoE module:** **(a)** Semantic states from SLUL generate query q to produce gating weights w_k over K pose experts. Each expert retrieves poses from a rule-based prior database, yielding the weighted blended pose $\mathbf{p}_{\text{pose}}$. **(b)** The blended poses are refined across temporal frames using smoothing loss $\mathcal{L}_{\text{smooth}}$, velocity loss $\mathcal{L}_{\text{vel}}$, and hand fidelity loss $\mathcal{L}_{\text{hand}}$ to ensure temporal coherence and spatial accuracy in the final stabilized sequence $\hat{P}_t$.

Pose-to-Video Conditioning. The stabilized pose sequence $\{\hat{P}_t\}$ drives a diffusion renderer (ControlNeXt/Wan) as control signals. The hierarchical objective coupling semantics, pose selection, and stability is

$$\mathcal{L} = \mathcal{L}_{\text{SLUL+SAGM}} + \lambda_{\text{MoE}}\mathcal{L}_{\text{MoE}} + \lambda_{\text{ent}}\mathcal{L}_{\text{ent}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{hand}}\mathcal{L}_{\text{hand}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}}.$$

At inference: encode $\mathbf{x}$, decode $\mathbf{g}$, gate experts to obtain $\mathbf{p}_{\text{pose}}$, stabilize into $\{\hat{P}_t\}$, then render video—training is end-to-end so semantic correctness and pose stability reinforce each other.

Pseudo-code (pose stage). As shown in Fig. 3, this stage turns gloss semantics into stable pose control signals: MoE gates specialize over pose priors, entropy keeps experts diverse, and smooth/hand/velocity penalties explicitly suppress jitter where sign intelligibility is most sensitive. The stabilized keypoints $\{\hat{P}_t\}$ then condition the diffusion renderer, so semantic correctness and motion stability are optimized jointly. Pseudo-code is given below.

Algorithm 2 SLP-MoE + stabilization

$q \leftarrow \text{Pool}(H)$; $w_k \leftarrow \text{softmax}(q^\top W_k)$
$\mathbf{p}_{\text{pose}} \leftarrow \sum_k w_k \phi_k(\mathbf{g})$
Refine $\mathbf{p}_{\text{pose}}$ into $\{\hat{P}_t\}$, extract $\hat{H}_t$
Compute $\mathcal{L}_{\text{MoE}}$, $\mathcal{L}_{\text{ent}}$, $\mathcal{L}_{\text{smooth}}$, $\mathcal{L}_{\text{hand}}$, $\mathcal{L}_{\text{vel}}$
$\mathcal{L} \leftarrow$ combine all terms

Overall, our methodology is not only innovative but also addresses some of the challenges that previous work failed to notice, making our approach both robust, efficient, and achieving a very high level of performance.

Type:	DEV SET BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE	TEST SET BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
NSLP-G (Hwang et al., 2021)	-	-	-	-	-	5.75	8.21	11.62	17.55	31.98
Fast-SLP Transformers (Fang et al., 2025c)	17.19	23.11	29.49	36.96	55.85	12.85	17.35	23.38	39.46	46.89
Neural Sign Actors (Baltatzis et al., 2024a)	-	-	-	-	-	13.12	18.25	25.44	41.31	47.55
SignLLM-1x1B-Super-P (ASL) (Fang et al., 2025a)	18.68	25.11	31.99	40.14	60.47	13.93	18.86	25.40	42.87	50.91
Stable Signer (Ours)	**23.24**	**30.41**	**39.14**	**47.85**	**70.68**	**15.67**	**21.65**	**28.49**	**48.87**	**59.56**

Table 1: **Comparison of the Pose Video generation performance of our model:** Compared with previous models, our parameters are comparable or even fewer, and we have the burden of semantic understanding, but our performance far exceeds the latest work. "-" indicates that the relevant data of the work has not been made public.

Approach:	DEV SET BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE	TEST SET BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
Base	15.09	21.98	31.07	59.40	62.09	12.99	16.07	26.82	55.32	58.93
SLUL	22.62	34.23	48.31	71.16	72.21	17.09	22.31	37.13	58.40	63.22
SLUL + SAGM Loss	23.24	30.41	39.14	47.85	70.68	15.67	21.65	28.49	48.87	59.56
SLUL + SAGM Loss + SLP MoE	**25.55**	**36.79**	**47.12**	**66.79**	**78.98**	**21.03**	**24.99**	**39.03**	**57.59**	**65.26**

Table 2: **Ablation Study:** A comparison of the generation performance under different settings of our model. The first three lines represent the generation of Pose Video, and the last line shows the final video generated by SLP MoE. We compared the impact of changes in different stages on the entire process, and found that our changes significantly improved the overall generation effect. **Base:** We use the base model designed by T5 (Raffel et al., 2020). **SLUL:** Sign Language Understanding Linker. **SLP-MoE:** Sign Language Production Mixture-of-Experts. **SAGM Loss:** Semantic-Aware Gloss Masking Loss. The results show that all the changes were beneficial.

3 Experiments

Here, we evaluate our Stable Signer, as well as SLUL, SLP MoE, SAGM Loss, etc. We conduct ablation evaluations, performance evaluations, efficiency evaluations, image quality evaluations, qualitative evaluations, and evaluations of prompt word meaning understanding, among others. We conduct a detailed and comprehensive test of our method. The experiments show that our method has significant improvements in all dimensions.

3.1 Experimental Setup

As mentioned above, our training is based on the Prompt2Sign dataset (Fang et al., 2025a) and the American Sign Language (ASL) portion of WLASL (Li et al., 2020). They have 30k and 10k video clips of ASL respectively, and some of the missing parts (Prompts and Glosses) were properly completed or deleted.

3.2 Evaluation for Sign Language Production

Back Translation of ASL. Back translation evaluates sign language production by translating generated videos back into text and comparing with original inputs using BLEU and ROUGE metrics. Higher scores indicate better semantic preservation and video quality. We evaluate Stable Signer on the How2Sign dataset and establish baselines for ASL video production, as shown in Table 1.

Approach:	DEV SET BLEU-4	ROUGE	TEST SET BLEU-4	ROUGE
Progressive Transformers (Saunders et al., 2020c)	15.18	49.46	15.92	46.57
Neural Sign Actors (Baltatzis et al., 2024a)	—	—	13.12	47.55
Fast-SLP Transformers (Fang et al., 2025c)	17.19	55.85	12.85	46.89
Stable Signer (Ours)	**25.55**	**78.98**	**21.03**	**65.26**
Δ *Acc.*	***+ 48.6%***	***+ 41.4%***	***+ 63.7%***	***+ 39.2%***

Table 3: **The performance comparison results of ASL Sign Video generation:** Unlike Table 1, this is the final comparison of the effects. The 3D Avatar is regarded as a qualified sign language video. The results show that our SLP MoE will produce better effects.

Table 1 shows that our method significantly outperforms existing approaches. We achieve BLEU-4 of 23.24 (dev) and 15.67 (test), substantially exceeding SignLLM's 18.68 and 13.93. Our model handles the complete pipeline including semantic understanding while achieving these improvements, making the results even more significant.

In Table 3, we present final video generation results with SLP-MoE. Compared to Fast-SLP Transformers (SignDiffusion (Fang et al., 2025c)), we achieve 48.6% improvement on BLEU-4 (dev) and 63.7% on test set. Our test BLEU-4 (21.03) surpasses latest Neural Sign Actors (13.12) (Baltatzis et al., 2024a) by a large margin, validating our hierarchical generation approach.

Ablation Study. We conduct ablation studies on the How2Sign dataset to validate each component's contribution, as shown in Table 2. We evaluate progressive improvements from base T5 (Raffel et al., 2020) through SLUL, SAGM Loss, and SLP-

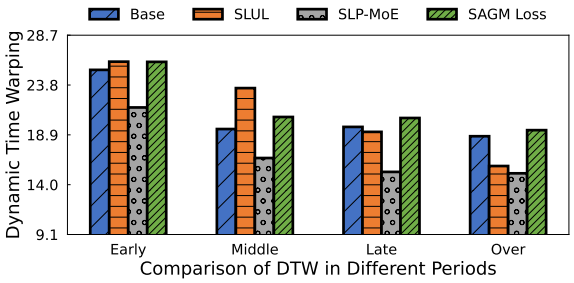


Figure 4: **Efficiency Study:** Comparing the DTW scores of different training periods (divided by 25% of an epoch for each training session), the lower the score, the better. We can observe that our SLUL and SLP MoE modifications have effectively improved the overall scores. The SAGM Loss, as it is a loss calculation and has no direct relation to performance, is a normal result. Therefore, we can say that all the modifications not only achieved our goals but also achieved at least a significant improvement in efficiency.

MoE additions on both development and test sets.

The base model achieves BLEU-4 of 15.09 (dev) and 12.99 (test). Adding SLUL significantly improves performance, with BLEU-3 jumping from 21.98 to 34.23 on dev set, demonstrating SLUL's effectiveness in translating prompts to glosses. Incorporating SAGM Loss adjusts the metrics with ROUGE reaching 70.68, reflecting its focus on semantic-aware masking. The complete system with SLP-MoE achieves best results: BLEU-4 of 25.55 (dev) and 21.03 (test), with ROUGE scores of 78.98 and 65.26. This demonstrates that semantic integration through MoE enables effective end-to-end learning for sign language production.

Training Efficiency Study. In Figure 4, we analyze training efficiency using DTW scores across epochs, partitioned into 25% intervals. Lower DTW values indicate better alignment between generated and ground truth sequences.

Both SLUL and SLP-MoE demonstrate faster convergence and better final performance. SLUL shows accelerated learning in early stages, suggesting effective semantic understanding provides stronger supervision. SLP-MoE achieves lowest DTW scores throughout, indicating the mixture-of-experts architecture enhances both final performance and training efficiency. SAGM Loss shows comparable curves to SLUL, as expected for a loss function modification.

Prompt Understanding Evaluation. We compare SLUL's Prompt2Gloss performance with previous Text2Gloss approaches in Table 4. This is important as users naturally input prompts like *"How do you sign 'I carried it'?"* rather than simplified text. Our approach achieves test BLEU-4 of 30.74 and ROUGE of 69.72, representing 43.3% and 23.5% improvements over SignLLM (Fang et al., 2025a). Despite handling more complex prompts than baseline Text2Gloss tasks, we outperform all methods including Zhang et al. (2025a) by 11.4% on test BLEU-4. This demonstrates that SAGM Loss effectively handles prompt nuances including ambiguity and implicit semantics. Consistent performance across dev and test sets indicates robust generalization rather than overfitting.

Approach:	DEV SET BLEU-4 ↑	DEV SET ROUGE ↑	TEST SET BLEU-4 ↑	TEST SET ROUGE ↑
Stoll *et al.*(Stoll et al., 2018)	16.34	48.42	15.26	48.10
Saunders (Saunders et al., 2020c)	20.23	55.41	19.10	54.55
Zhang (Zhang et al., 2025a)	-	-	27.60	66.40
SignLLM (Fang et al., 2025a)	23.10	58.76	22.05	56.46
Stable Signer (Ours)	**32.08**	**76.54**	**30.74**	**69.72**
Δ *Acc.*	***+ 38.9%***	***+ 30.2%***	***+ 43.3%***	***+ 23.5%***

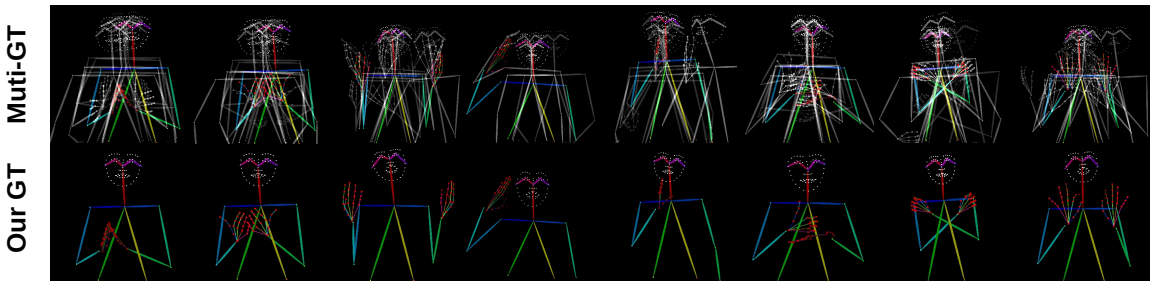
Table 4: **Prompt Accuracy:** Due to the scarcity of similar work to ours, we compared our semantic understanding task with the similar Text2Gloss task. Although our task was more challenging, we still outperformed previous works, as well as those in the same period.

3.3 Evaluation for Video Generation Quality

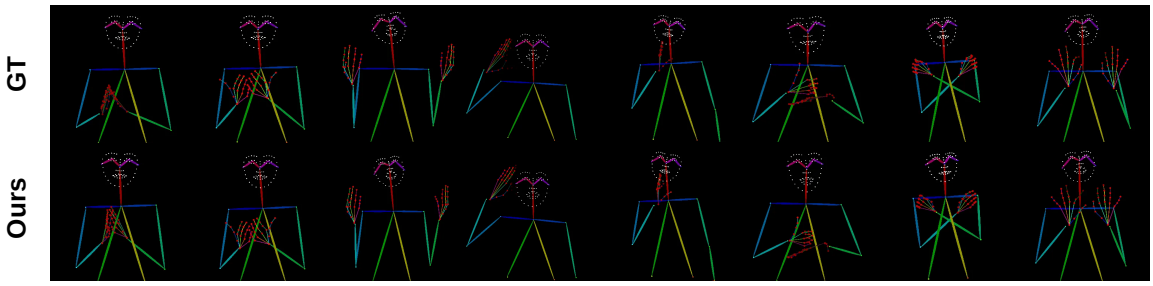
We evaluate video generation quality through both qualitative visualizations and quantitative metrics, comprehensively assessing how our SLP-MoE module converts pose sequences into realistic, high-quality sign language videos that accurately convey semantic meaning.

Qualitative Evaluations. Figure 5 provides a comprehensive visualization of our complete end-to-end sign language production pipeline across three progressive stages: (a) simplified pose representations automatically learned by the model, (b) intermediate pose video frames generated during the conversion process, and (c) final rendered sign language videos presented alongside ground truth frames for direct comparison.

The simplified pose targets shown in Figure 5(a) demonstrate a critical innovation of our approach—the model automatically learns to generate clean, semantically meaningful pose representations through the synergistic integration of rule-based pose priors and SLUL-based semantic guidance. These simplified poses effectively eliminate the noise, jitter, and temporal instability commonly present in raw pose estimations from conventional video analysis tools such as DWPose (Yang et al., 2023). Of course, what is even more important



(a) The goal of learning vs. the poses that exist in the dataset



(b) Text2Pose model predicted pose video vs. GT pose video

(c) The various styles of Sign Videos that we ultimately produced

Figure 5: **Qualitative Results & User Study.** We visualize our end-to-end sign language production pipeline. (a) Simplified pose targets automatically learned by the model, enabling robust end-to-end learning for SLP—an approach increasingly adopted in recent studies. (b) Intermediate pose video frames generated during the process. (c) Final sign language video frames. Ground truth frames are provided for comparison. Our method achieves high-fidelity image generation while accurately conveying sign language pose information.

is as shown in the first row of the picture: different sign language users have different expressions. These gestures are not aligned in time and space, which causes the models that previously learned average gestures to often be unstable, or to generate restricted or unconfident gestures. However, our method will automatically learn deterministic gestures. As long as one of the multiple gestures is completed, it is considered correct. Therefore, our generation part is more stable.

By establishing this clean pose foundation, our approach enables more robust and reliable end-to-end training—a paradigm increasingly adopted in recent sign language production research. The clean representations ensure that semantic information from SLUL is properly conveyed without being corrupted by pose estimation artifacts. This part of the assessment also indicates that our new conceptual framework has been successful.

The intermediate pose videos presented in Figure 5(b) reveal the quality and effectiveness of our Gloss2Pose conversion process. These frames showcase smooth temporal transitions between different sign gestures, accurate gesture trajectories that properly capture the dynamic nature of sign language movements, and appropriate timing that maintains the natural rhythm of signing. This approach enables the model to automatically learn

	SSIM ↑	Hand SSIM ↑	Hand Pose ↓	FID ↓
EDN (Chan et al., 2019b)	0.737	0.553	23.09	41.54
vid2vid (Wang et al., 2018a)	0.750	0.570	22.51	56.17
Pix2PixHD (Wang et al., 2018b)	0.737	0.553	23.06	42.57
Stoll et al. (Stoll et al., 2020b)	0.727	0.533	23.17	64.01
SIGNGAN (Saunders et al., 2022)	0.759	0.605	22.05	27.75
SignDiff (Fang et al., 2025c)	0.849	0.676	20.04	25.22
Stable Signer (Ours)	**0.892**	**0.732**	**17.68**	**21.04**

Table 5: **Video quality study:** Compared with existing and the latest hand gesture image production work, ours is the most performant, and it far exceeds the mainstream work. This is precisely why the videos we generate have better smoothness and fewer errors.

	SSIM ↑	Hand SSIM ↑	Hand Pose ↓	FID ↓
Baseline (Saunders et al., 2022)	0.743	0.582	22.87	39.33
Hand Discriminator	0.738	0.565	22.81	39.22
Hand Keypoint Loss	0.759	0.605	22.05	27.75
Zhang (Zhang et al., 2025a)	0.864	-	-	56.28
SignDiff (Fang et al., 2025c)	0.849	0.676	20.04	25.22
Stable Signer (No SLP-MoE) (Ours)	**0.872**	**0.716**	**19.24**	**24.22**
Stable Signer (Ours)	**0.892**	**0.732**	**17.68**	**21.04**

Table 6: **Ablation Study of SLP-MoE's Video Quality:** We also attempted to conduct an ablation study on our SLP-MoE to verify whether it was the improvement of the base model or our method that was responsible for the effect. The results showed that although the better backbone played a role, our method could achieve even greater improvements.

the appropriate GT, making our generation process more robust and laying the foundation for the end-to-end learning of the entire complex system.

Most importantly, the final generated sign language videos shown in Figure 5(c) achieve high visual fidelity and photorealistic quality that closely matches authentic human signing. When directly compared with ground truth video frames displayed alongside, our generated videos successfully maintain semantic accuracy while producing smooth, natural-looking sign language gestures.

Baseline Comparison. We conduct comprehensive quantitative comparisons with SOTA sign language video generation methods using four complementary evaluation metrics (SSIM&Hand SSIM (Wang et al., 2004), Hand Pose error measuring (Ge et al., 2019), FID (Heusel et al., 2017)) that capture different aspects of generation quality. Table 5 shows Stable Signer achieves superior performance across all metrics, our image quality is also far superior to that of SOTA or the latest works.

Ablation Study on Video Generation. To isolate SLP-MoE's contribution and verify improvements come from methodology rather than just better backbones, we conduct ablations in Table 6.

Results show clear progressive improvements. Our "No SLP-MoE" variant achieves 0.872 SSIM and 24.22 FID, outperforming SignDiff and validating that our simplified pipeline and SLUL already provide benefits. Full Stable Signer with SLP-MoE reaches 0.892 SSIM and 21.04 FID, demonstrating MoE adds significant value. Hand Pose improves from 19.24 to 17.68 (8.1%), showing semantic integration helps generate more accurate positions critical for intelligibility.

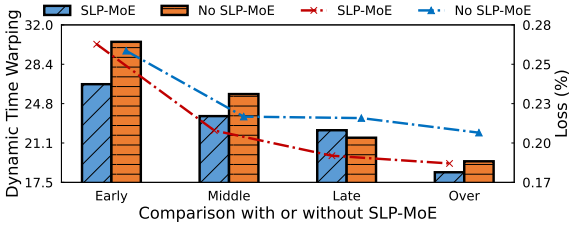


Figure 6: **Ablation Study on SLP-MoE's Training Efficiency:** We examined the efficiency under the condition where there was no SLP-MoE (using the regular Pose2Vid model instead (Cheng et al., 2025)). The results showed that using our SLP-MoE for joint training could achieve more thorough convergence, train faster, and produce better results.

Figure 6 shows SLP-MoE achieves better final performance and faster convergence. DTW scores decrease more rapidly with MoE, suggesting semantic information provides stronger learning signals. The performance gap remains substantial at convergence, indicating MoE enables better optimization. Loss curves show comparable convergence rates, confirming gains come from enhanced capability to leverage semantics rather than easier optimization. This validates our hypothesis that end-to-end joint training of semantic understanding and generation yields superior results.

4 Conclusion

In this paper, we observed that the current goals of the sign language generation tasks have fallen into suboptimal average goals due to multiple Pose Ground Truth. Therefore, we attempted to shorten the production chain and developed the Stable Signer framework for end-to-end SLP, which includes SLUL for complex semantic understanding, SLP-MoE for generating high-quality sign language videos, and SAGM Loss for end-to-end training. The results show that, in the ASL production process, with the same dataset, it can achieve a significant improvement of approximately 50% in the accuracy of poses compared to the existing traditional methods. Moreover, it has achieved results that surpass the state-of-the-art in nearly ten comprehensive and rich experiments.

Limitations

There might be a slight flickering in the transition between different poses of the task, but our SLP MoEs will eliminate this problem. According to the qualitative assessment, this does not affect the final outcome.

References

2023. Introducing Falcon LLM . https://falconllm.tii.ae.

Sharif Md. Abdullah, Abhijit Paul, Shebuti Rayana, Ahmedul Kabir, and Zarif Masud. 2025. State-of-the-art translation of text-to-gloss using mbart : A case study of bangla. *Preprint*, arXiv:2504.02293.

Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. 2024a. 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 (CVPR)*.

Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. 2024b. Neural sign actors: A diffusion model for 3d sign language production from text. *Preprint*, arXiv:2312.02702.

Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. 2019a. Everybody dance now. *Preprint*, arXiv:1808.07371.

Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. 2019b. Everybody Dance Now. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*.

Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, and 1 others. 2025. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*.

Hyunsoo Cho, Hyuhng Joon Kim, Junyeob Kim, Sang-Woo Lee, Sang-goo Lee, Kang Min Yoo, and Taeuk Kim. 2022. Prompt-augmented linear probing: Scaling beyond the limit of few-shot in-context learners. *CoRR*, abs/2212.10873.

Sen Fang, Chen Chen, Lei Wang, Ce Zheng, Chunyu Sui, and Yapeng Tian. 2025a. Signllm: Sign language production large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 6622–6634.

Sen Fang, Chunyu Sui, Hongwei Yi, Carol Neidle, and Dimitris N. Metaxas. 2025b. Signx: The foundation model for sign recognition. *Preprint*, arXiv:2504.16315.

Sen Fang, Chunyu Sui, Yanghao Zhou, Xuedong Zhang, Hongbin Zhong, Yapeng Tian, and Chen Chen. 2025c. Signdiff: Diffusion model for american sign language production. In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–11.

Sen Fang, Hongbin Zhong, Yalin Feng, and Dimitris N. Metaxas. 2025d. Streamflow: Theory, algorithm, and implementation for high-efficiency rectified flow generation. *Preprint*, arXiv:2511.22009.

Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 2019. 3D Hand Shape and Pose Estimation from a Single RGB Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*.

Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. 2023. Continuous sign language recognition with correlation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Eui Jun Hwang, Jung-Ho Kim, and Jong C. Park. 2021. Non-autoregressive sign language production with gaussian space. In *The 32nd British Machine Vision Conference (BMVC 21)*. British Machine Vision Conference (BMVC).

Mert Inan, Anthony Sicilia, and Malihe Alikhani. 2025. SignAlignLM: Integrating multimodal sign language processing into large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3691–3706, Vienna, Austria. Association for Computational Linguistics.

Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. SignCLIP: Connecting text and sign language by contrastive learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9171–9193, Miami, Florida, USA. Association for Computational Linguistics.

Michael Kipp, Alexis Heloir, and Quan Nguyen. 2011. Sign language avatars: Animation and comprehensibility. In *International Workshop on Intelligent Virtual Agents*, pages 113–126. Springer.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469.

Xiao Liu, Guangyi Chen, Yansong Tang, Guangrun Wang, Xiao-Ping Zhang, and Ser-Nam Lim. 2024. Language-free compositional action generation via decoupling refinement. *Preprint*, arXiv:2307.03538.

Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Ying Shan, Xiu Li, and Qifeng Chen. 2024. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *Preprint*, arXiv:2304.01186.

Amit Moryossef, Gerard Sant, and Zifan Jiang. 2025. Pose-based sign language appearance transfer. *Preprint*, arXiv:2410.13675.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. 2025. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *Preprint*, arXiv:2407.02371.

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

Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. 2025. Controlnext: Powerful and efficient control for image and video generation. *Preprint*, arXiv:2408.06070.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. 2021. Scaling vision with sparse mixture of experts. *Preprint*, arXiv:2106.05974.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pages 36479–36494. Curran Associates, Inc.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020a. Everybody sign now: Translating spoken language to photo realistic sign language video. *Preprint*, arXiv:2011.09846.

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 Camgöz, and Richard Bowden. 2020c. Progressive Transformers for End-to-End Sign Language Production. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

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

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

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021c. 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.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2022. 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 (CVPR)*, pages 5141–5151.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *Preprint*, arXiv:1701.06538.

Gregory Shuler, Lisa Mistler, Kathleen Torrey, and Rayne depukat. 2013. Bridging communication gaps with the deaf. *Nursing*, 43.

Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2018. Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2020a. Text2Sign: Towards Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. *International Journal of Computer Vision (IJCV)*.

Stephanie Stoll, Simon Hadfield, and Richard Bowden. 2020b. Signsynth: Data-driven sign language video generation. In *European Conference on Computer Vision*, pages 353–370. Springer.

Haowen Sun, Ruikun Zheng, Haibin Huang, Chongyang Ma, Hui Huang, and Ruizhen Hu. 2024. Lgtm: Local-to-global text-driven human motion diffusion model. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, SIGGRAPH '24, page 1–9. ACM.

Laia Tarrés, Gerard I. Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró i Nieto. 2023. Sign language translation from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. 2022. Human motion diffusion model. *Preprint*, arXiv:2209.14916.

Špela Vintar, Boštjan Jerko, and Marjetka Kulovec. 2012. Compiling the slovene sign language corpus. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. Language Resources and Evaluation Conference (LREC)*, volume 5, pages 159–162.

Harry Walsh, Ed Fish, Ozge Mercanoglu Sincan, Mohamed Ilyes Lakhal, Richard Bowden, Neil Fox, Bencie Woll, Kepeng Wu, Zecheng Li, Weichao Zhao, Haodong Wang, Wengang Zhou, Houqiang Li, Shengeng Tang, Jiayi He, Xu Wang, Ruobei Zhang, Yaxiong Wang, Lechao Cheng, and 3 others. 2025. Slrtp2025 sign language production challenge: Methodology, results, and future work. *Preprint*, arXiv:2508.06951.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018a. Video-to-Video Synthesis. In *Advances in Neural Information Processing Systems (NIPS)*.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018b. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*.

Pan Xie, Qipeng Zhang, Taiyi Peng, Hao Tang, Yao Du, and Zexian Li. 2023. G2p-ddm: Generating sign pose sequence from gloss sequence with discrete diffusion model. *Preprint*, arXiv:2208.09141.

Frank F. Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. 2022. A systematic evaluation of large language models of code. In *MAPS@PLDI*.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. 2023. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220.

Aoxiong Yin, Haoyuan Li, Kai Shen, Siliang Tang, and Yueting Zhuang. 2024. T2s-gpt: Dynamic vector quantization for autoregressive sign language production from text. *Preprint*, arXiv:2406.07119.

Jan Zelinka and Jakub Kanis. 2020. Neural sign language synthesis: Words are our glosses. pages 3384–3392.

Han Zhang, Rotem Shalev-Arkushin, Vasileios Baltatzis, Connor Gillis, Gierad Laput, Raja Kushalnagar, Lorna C Quandt, Leah Findlater, Abdelkareem Bedri, and Colin Lea. 2025a. Towards ai-driven sign language generation with non-manual markers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023a. Adding conditional control to text-to-image diffusion models. *Preprint*, arXiv:2302.05543.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2022. Motiondiffuse: Text-driven human motion generation with diffusion model. *Preprint*, arXiv:2208.15001.

Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. 2023b. Controlvideo: Training-free controllable text-to-video generation. *Preprint*, arXiv:2305.13077.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. 2025b. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *Preprint*, arXiv:2406.19680.

Lei Zhong, Yi Yang, and Changjian Li. 2025. Smoogpt: Stylized motion generation using large language models. *Preprint*, arXiv:2509.04058.

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

A Related Works

A.1 Sign Language Production

Sign Language Production (SLP) aims to generate sign language videos or motion sequences from spoken language inputs, bridging communication gaps for the deaf and hard-of-hearing community (Shuler et al., 2013). Early deep learning approaches decomposed SLP into cascaded steps: Text-to-Gloss (T2G) (Abdullah et al., 2025), Gloss-to-Pose (G2P) (Xie et al., 2023), and Pose-to-Sign (P2S) (Moryossef et al., 2025). However, these pipeline methods struggled with smooth motion blending between signs. Recent works have explored end-to-end approaches using transformer architectures and diffusion models. Neural Sign Actors (Baltatzis et al., 2024b) proposed a diffusion-based model trained on 4D signing avatars with anatomically informed graph neural networks defined on the SMPL-X skeleton for realistic 3D sign production. T2S-GPT (Yin et al., 2024) introduced dynamic vector quantization for autoregressive SLP, while SignLLM (Fang et al., 2025a) extended SLP to a multilingual framework supporting eight sign languages through novel reinforcement learning components. Despite these advances, generating realistic, semantically accurate signing motions with fine-grained hand and facial expressions remains challenging (Walsh et al., 2025; TII, 2023; Cho et al., 2022; Fang et al., 2025b; Hwang et al., 2021).

A.2 Motion Generation

Human motion generation has witnessed significant progress with the adoption of diffusion models. MotionDiffuse (Zhang et al., 2022) was among the first diffusion-based frameworks for text-driven motion generation, demonstrating probabilistic mapping through denoising steps and enabling fine-grained body part control. MDM (Motion Diffusion Model) (Tevet et al., 2022) introduced a transformer-based architecture with classifier-free guidance, predicting motion samples rather than noise to facilitate geometric losses on joint positions and velocities. LGTM (Sun et al., 2024) proposed a Local-to-Global pipeline leveraging large language models to decompose motion descriptions into part-specific narratives, addressing local semantic accuracy issues. Recent works have also explored music-to-dance synthesis, compositional action generation (Liu et al., 2024), and stylized motion control (Zhong et al., 2025). These ad-

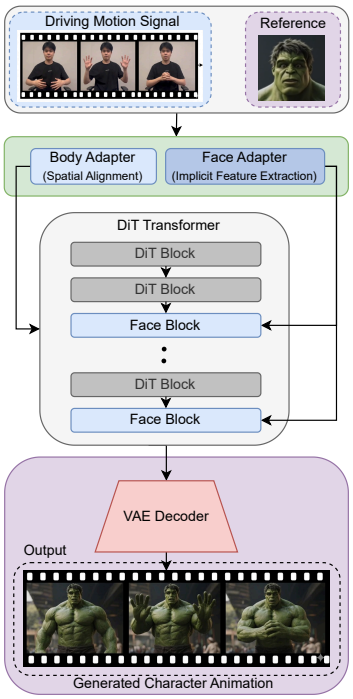


Figure 7: The framework of Wan-Animate (Cheng et al., 2025). It utilizes a unified input paradigm and a decoupled control strategy. Spatially-aligned skeleton signals are directly injected into the DiT backbone for body control, while facial expressions are driven by implicit features extracted from reference images and injected via cross-attention layers.

vances provide foundational techniques for generating temporally coherent human movements, which are directly applicable to sign language motion synthesis where precise hand articulation and body coordination are essential.

A.3 Controllable Video Generation

Controllable video generation has emerged as an active research direction, extending image-based control mechanisms to the temporal domain. ControlNet (Zhang et al., 2023a) pioneered the integration of structural controls such as depth maps, edge detection, and human pose into text-to-image diffusion models through zero convolution layers. ControlVideo (Zhang et al., 2023b) adapted ControlNet to video generation without fine-tuning, inheriting high-quality consistent frame generation through fully cross-frame interaction and interleaved-frame smoothing. ControlNeXt (Peng et al., 2025) further improved efficiency by reducing trainable parameters up to 90% while supporting diverse conditional controls including pose sequences for video diffusion. Follow-Your-Pose (Ma et al., 2024) and MimicMotion (Zhang et al., 2025b) specifically addressed human video generation with pose guidance. Most recently, Wan-Animate (Cheng et al., 2025) advances this field by leveraging a Diffusion Transformer (DiT) (Peebles and Xie, 2023; Tevet et al., 2022; Fang et al., 2025d; Saharia et al., 2022) architecture. As illustrated in Figure 7, it employs a decoupled control paradigm that integrates spatially-aligned skeleton signals for body movements and implicit facial features via cross-attention for fine-grained expression reenactment. These controllable generation techniques offer promising directions for sign language video synthesis (Kipp et al., 2011; Vintar et al., 2012; Xu et al., 2022), where pose-conditioned generation can ensure anatomically correct and semantically aligned signing motions.

B Discussion

In this paper, we conducted thorough and comprehensive experiments: for various ASL generation effects, we compared the current latest sign language work under fair conditions, and we achieved a significant improvement over them. For ablation assessment, due to the sequential dependencies of various components, we gradually added to verify that our modifications were effective (all components, including the a priori component, were evaluated). For efficiency assessment and efficiency metric ablation assessment, we conducted comprehensive experiments and comparative experiments. Qualitative assessment and image quality assessment clearly demonstrated the superiority of our model quality and the advantages of our method.

But, we also noticed that some recent generation works such as SignCLIP (Jiang et al., 2024), SignAlignLLM (Inan et al., 2025), Sign as Token (Zuo et al., 2025), etc., share similar goals with ours but are different. Moreover, they use different datasets for training. Therefore, we cannot easily compare them. Up to now, our work on ASL SLP should be the best: many sign language works cannot be directly compared with ours due to minor differences in goals. Finally, the non-ASL production is beyond our scope of work. We can consider conducting research on this in the future.