SEMOSIGN: SEMANTIC-MOTION-GUIDED FLOW MATCHING FOR SIGN LANGUAGE PRODUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sign language production (SLP) has adopted flow-matching models to synthesize sign motions from spoken-language input. However, existing approaches typically rely on learned flow trajectories without explicitly coordinating semantic representation learning with motion-validity guidance during sampling. This may lead to insufficient language–motion correspondence, trajectory drift, and semantically inconsistent or implausible motions. To address these limitations, we propose *SeMoSign*, a unified semantic-motion-guided flow-matching framework that promotes consistent generation from latent semantic grounding to trajectory evolution. During training, Semantic-Conditioned Latent Modulation (SCLM) injects textual semantic priors into latent motion representations, establishing structured language–motion correspondence without altering the flow-matching backbone. During inference, Per-step Energy Guidance (PEG) evaluates provisional sampling states with a learned region-aware motion-energy landscape and applies differentiable corrections toward plausible motion configurations at selected steps. By coupling semantic latent guidance with motion-aware trajectory refinement, SeMoSign maintains semantic fidelity while improving the coherence and plausibility of generated motions. Experiments on PHOENIX-2014T demonstrate that SeMoSign improves semantic- and motion-related metrics over representative flow-matching-based SLP methods. Qualitative results further show enhanced semantic consistency, temporal coherence, and motion plausibility.

1 INTRODUCTION

Sign language production (SLP) aims to automatically generate continuous sign motion sequences from spoken-language text Camgöz et al. (2017). We focus on text-to-pose SLP, where the input is spoken-language text and the output is a temporally ordered sequence of 3D body, hand, and facial poses. Unlike gloss-to-pose generation, sign language translation, and text-to-video synthesis, this task directly generates pose sequences from spoken-language input without requiring a predefined gloss sequence or producing RGB videos. The generated motions should preserve the linguistic content while maintaining temporal continuity and physical plausibility. Early studies explored statistical translation and template-based animation for SLP, but their motions were often limited in continuity and realism Bragg et al. (2019); Camgoz et al. (2018); Hwang et al. (2021); McDonald et al. (2015); Othman & Jemni (2011); Zelinka & Kanis (2020); Camgöz et al. (2020); Huang et al. (2015); Fang et al. (2017). Deep learning subsequently enabled direct text-to-motion modeling, with Transformer-based systems improving the temporal coherence of generated signs Stoll et al. (2020); Saunders et al. (2020a; 2021a); Tang et al. (2022a); Viegas et al. (2023); Vaswani et al. (2017); Yin (2020); Lipman et al. (2022). Other studies introduced autoregressive prediction, adversarial learning, mixture-density modeling, and denoising-based diffusion processes to improve motion diversity, expressiveness, and temporal consistency Saunders et al. (2020a;b); Stoll et al. (2020); Tang et al. (2022a); Ho et al. (2020). More recently, diffusion-based and flow-based generative models have become important approaches for text-to-pose SLP. Diffusion models generate motions through iterative denoising, whereas flow-matching models learn a continuous velocity field that transports noise to motion latents Ho & Salimans (2022); Lipman et al. (2022). These models have further improved the realism, diversity, and temporal coherence of generated signs Lipman et al. (2022); Rombach et al. (2022); Zhang et al. (2024); Petrovich et al. (2022). Representative methods, including G2P-DDM Xie et al. (2024), Sign-IDD Tang et al. (2025a), Gloss-Driven Conditional Diffusion

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

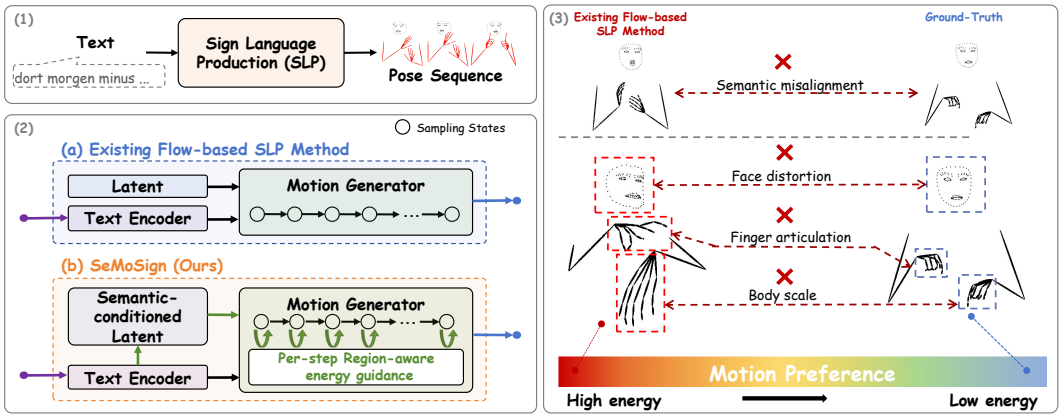

Figure 1: (1) A sign language production model generates a pose sequence from a text prompt. (2) Existing flow-based methods lack explicit semantic and learned motion-energy guidances. Our method introduces semantic-conditioned latent representations through SCLM and applies per-step, region-aware energy guidance through PEG. (3) Comparison with ground-truth motion reveals semantic misalignment and motion-quality issues in existing flow-based SLP methods.

Models Tang et al. (2025b), Neural Sign Actors Baltatzis et al. (2023), and SignDiff Fang et al. (2025), formulate sign motion generation as an iterative process and substantially improve over previous deterministic approaches. However, existing flow-based formulations focus on learning the data distribution or improving sampling efficiency, without explicitly evaluating or correcting whether intermediate states remain in regions associated with plausible sign motions Lipman et al. (2022); Xie et al. (2024); Tang et al. (2025a;b); Baltatzis et al. (2023); Fang et al. (2025). The trajectory may therefore drift toward invalid motion regions, causing semantic inconsistencies, temporal instability, or physically implausible movements. This challenge has two aspects: the latent motion representation may be insufficiently structured by linguistic semantics, while the sampling trajectory may lack an explicit motion-validity prior. Realistic SLP requires both a semantically structured latent representation and constraints on trajectory evolution.

Existing SLP methods introduce language conditions through textual encoders, conditional feature integration, or language–motion interaction mechanisms Stoll et al. (2020); Saunders et al. (2020a;b); Tang et al. (2022b; 2025b); Viegas et al. (2023). Although these approaches allow textual information to influence motion generation, semantic information is injected after forming the latent motion representation Zhou et al. (2023). The latent motion space is not explicitly organized according to the input sentence, limiting precise semantic–motion correspondence. Directly incorporating sentence-level semantic priors into the latent motion representation remains underexplored in SLP. This motivates a lightweight semantic-aware mechanism that introduces linguistic information directly into the latent space without substantially modifying the generation backbone.

The sampling process also lacks constraints on motion validity. Existing flow-based SLP methods iteratively integrate the learned velocity field without evaluating whether intermediate states remain within realistic sign motion regions. Once intermediate states move toward low-plausibility motion regions, later steps may fail to recover the accumulated errors, leading to unrealistic joint configurations, unstable hand trajectories, or inconsistent coordination among the hands, face, and body. Additional physical constraints during training or post-hoc refinement have been explored Karunratanakul et al. (2023); Ho & Salimans (2022); Du et al. (2023); Zhang et al. (2025b). However, training-time constraints may complicate optimization, while post-hoc refinement cannot correct errors arising at intermediate states. An effective mechanism that provides differentiable corrections to intermediate states and guides the trajectory toward realistic motion regions is therefore desirable.

To address these issues, we propose SeMoSign, a semantic-motion-guided flow-matching framework for sign language production. Specifically, Semantic-Conditioned Latent Modulation (SCLM) injects sentence-level semantic priors directly into the latent motion representation during flow-matching training, promoting closer linguistic–motion correspondence while preserving the original flow-matching formulation. We further propose Per-step Energy Guidance (PEG), which learns a

region-aware motion-energy landscape based on facial motion, left- and right-hand articulation, and coordination among these regions. PEG evaluates provisional sampling states and applies differentiable corrections at selected steps. Thus, SCLM provides semantic structure for the latent representation, while PEG constrains motion validity as the trajectory evolves. Together, they reduce trajectory drift and improve semantic consistency and physical plausibility. Extensive experiments on the PHOENIX-2014T benchmark show that SeMoSign improves both semantic- and motion-level evaluation metrics over representative SLP baselines. Qualitative comparisons on PHOENIX-2014T and USTC-CSL further show more coherent and faithful sign motions, with fewer implausible configurations and more stable trajectories. Our contributions are summarized as follows:

- We propose a semantic-aware latent representation learning strategy that incorporates sentence-level semantic priors into latent motion space, improving semantic–motion correspondence and providing a structured foundation for flow-matching-based SLP.
- We introduce an energy-guided refinement strategy that learns a region-aware differentiable motion-energy landscape and applies sampling corrections, guiding trajectories toward plausible regions while reducing drift and anatomically unrealistic motions.
- We conduct quantitative and qualitative experiments on the PHOENIX-2014T and USTC-CSL benchmarks. The results and ablation studies provide empirical evidence that the proposed framework improves semantic consistency and motion plausibility.

2 Related Work

2.1 Sign Language Production

Sign language production (SLP) aims to generate continuous sign motions from spoken-language text or glosses Bragg et al. (2019); Camgoz et al. (2018); Koller et al. (2017); Hu et al. (2021); Joze & Koller (2018). Studies consider gloss-to-pose and text-to-pose generation; this work focuses on 3D pose generation from spoken-language text. Early systems decomposed SLP into text-to-gloss translation and motion synthesis, relying on symbolic representations or predefined templates. Such designs limited co-articulation and natural temporal dynamics Camgoz et al. (2018); Hwang et al. (2021); McDonald et al. (2015); Othman & Jemni (2011); Stoll et al. (2020); Zelinka & Kanis (2020); Koller et al. (2019); Bowden (2016). Transformer-based methods formulated SLP as conditional sequence generation and incorporated adversarial training, mixture density modeling, gloss supervision, graph representations, and motion priors to improve expressiveness and temporal consistency Stoll et al. (2020); Saunders et al. (2020a; 2021a); Vaswani et al. (2017); Tang et al. (2022b); He et al. (2025); Saunders et al. (2020b; 2021b; 2022).

Diffusion-based generative models have recently become an important paradigm for SLP Song et al. (2020). These methods generate sign motions through iterative denoising and have improved motion realism, diversity, and temporal coherence Baltatzis et al. (2023); Xie et al. (2024); Tang et al. (2025b;a); Fang et al. (2025). Flow matching provides a related continuous generation paradigm that learns a velocity field to transport noise toward the data distribution, rather than explicitly modeling a diffusion denoising process Lipman et al. (2022). Its deterministic transport formulation offers a flexible alternative for latent motion generation and sampling. However, existing flow-based formulations focus on learning the data distribution or improving sampling efficiency, without explicitly constraining intermediate states to remain on the manifold of realistic sign motions. Their generation trajectories may remain susceptible to semantic ambiguity and accumulated sampling errors.

2.2 Semantic-aware Representation Learning

Semantically meaningful motion representations are essential for SLP, since subtle linguistic differences may correspond to distinct body and hand movements Camgoz et al. (2018). Existing methods inject semantic information through conditional encoders, gloss representations, classifier-free guidance, or cross-modal attention Stoll et al. (2020); Tang et al. (2022b); Ho & Salimans (2022); Raffel et al. (2020); Jiao et al. (2024), improving text–motion alignment. Nevertheless, semantic information is often injected after the latent motion representation is formed. The latent space is weakly constrained by semantic priors, limiting structured semantic–motion correspondence. Explicitly incorporating priors into latent representation learning remains underexplored in flow-based SLP.

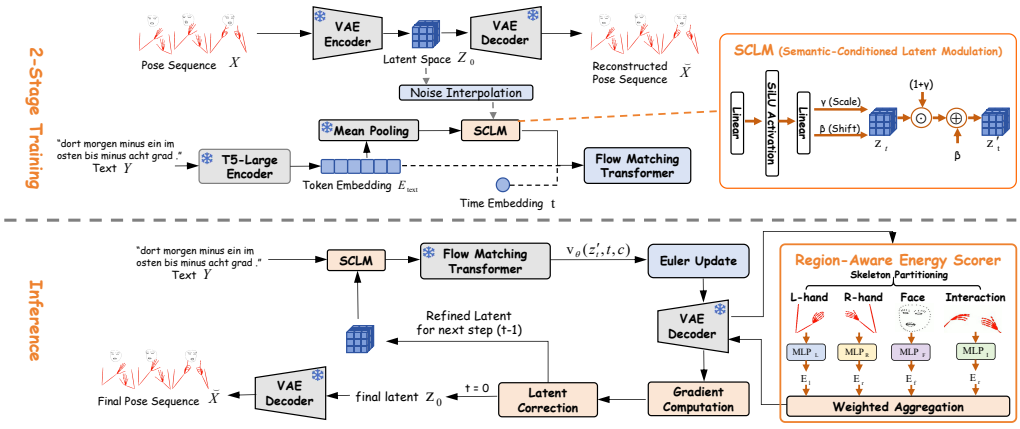

Figure 2: Overview of the proposed SeMoSign framework. SeMoSign follows a sequential semantic-to-motion generation process. Semantic-Conditioned Latent Modulation (SCLM) injects sentence-level semantic information into the latent state before flow-velocity prediction. The flow field produces a provisional sampling state, which is evaluated by Per-step Energy Guidance (PEG) during inference. PEG applies differentiable motion-energy corrections to guide state evolution.

2.3 Energy-Guided Generative Modeling

Energy-based models learn landscapes that distinguish realistic from unrealistic samples and guide states toward low-energy regions Zhang et al. (2025a;b); Du et al. (2023). Recent studies integrate energy guidance into models for image generation, restoration, and human motion synthesis by refining sampling states Karunratanakul et al. (2023); Du et al. (2023); Zhang et al. (2025b). Despite these advances, energy-guided sampling remains largely unexplored in SLP Tang et al. (2022b). Existing SLP methods rely on learned generation fields without explicitly correcting intermediate states. Once trajectories deviate from the realistic sign-motion manifold, accumulated errors may degrade motion quality and semantic consistency. This motivates inference-time guidance that provides differentiable corrections toward realistic sign-motion regions Zhang et al. (2025b).

3 Methodology

Given a sentence, SeMoSign generates a sign motion sequence using a latent flow-matching model Lipman et al. (2022); Rombach et al. (2022). As illustrated in Figure 2, during training, the ground-truth pose sequence X is encoded into the latent space by a pretrained VAE encoder and interpolated with a noise latent to obtain the flow-matching state z_t. The text prompt Y is processed by a pretrained T5-Large encoder to extract semantic features. Their mean-pooled representation is fed into the Semantic-Conditioned Latent Modulation (SCLM) module to modulate z_t, yielding the semantic-conditioned state $\tilde{z}_t$, which is then provided to the flow-matching Transformer together with the textual condition and time embedding to predict the conditional velocity. During inference, the process starts from a noise latent. At each sampling step, SCLM first modulates the latent, and the flow-matching Transformer followed by a reverse Euler update produces a provisional latent state. This state is decoded into motion space and evaluated by the Per-step Energy Guidance (PEG) module, which models facial motion, left- and right-hand motion, and inter-region coordination to compute a region-aware motion energy. The energy gradient is then used to correct the provisional state, and the refined latent is propagated to the next sampling step. Finally, the latent is decoded by the VAE decoder to generate the sign pose sequence. SCLM is optimized jointly with the flow-matching Transformer, whereas PEG is trained independently and applied during inference.

3.1 Semantic-Conditioned Latent Modulation

Existing flow-based sign language production methods introduce textual conditions through cross-attention, feature concatenation, or conditional feature injection within the velocity network. However, these strategies mainly affect velocity prediction after the latent state has been formed, leaving

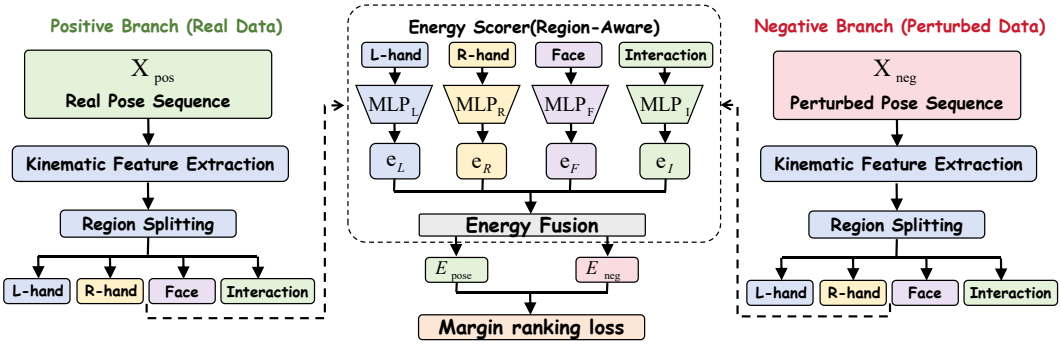

Figure 3: Learning the motion energy landscape. Real and corrupted sign motion sequences are decomposed into four motion components and processed by regional energy estimators. Their energies are aggregated and optimized with a margin-based ranking loss, assigning lower energies to motions. The differentiable landscape provides gradient guidance during flow-matching sampling.

the latent motion representation weakly structured by sentence-level semantics. To address this limitation, we propose Semantic-Conditioned Latent Modulation (SCLM), which adapts the intermediate latent state according to the sentence before it is processed by the flow predictor. As illustrated in the training branch of Figure 2, SCLM is inserted between Noise Interpolation and the Flow Matching Transformer, while the same modulation is applied before velocity prediction at each inference step. In this way, SCLM provides a sentence-aware latent foundation for flow-based generation without modifying the underlying flow-matching backbone.

Let X denote a ground-truth sign pose sequence and let E_ϕ denote a pretrained and frozen VAE encoder Kingma & Welling (2013). The motion latent is $z_{\text{data}} = E_\phi(X)$, corresponding to Z_0 in Figure 2, and z_t denotes the intermediate latent obtained from the noise interpolation process. Given a text prompt y, the frozen T5-Large encoder produces token-level embeddings $E_{\text{text}} = \text{TextEnc}(y)$, which are provided to the Flow Matching Transformer as the textual condition. In parallel, the Mean Pooling module in the SCLM block aggregates E_{text} into a sentence-level representation. This representation is processed by the trainable MLP controller M_ψ, consisting of linear layers and a SiLU activation, to predict the channel-wise scale and shift parameters:

$$[\gamma_y, \beta_y] = M_\psi\left(\text{MeanPool}(E_{\text{text}})\right). \tag{1}$$

The parameters are broadcast along the temporal dimension and used to modulate the intermediate latent through the residual affine transformation shown in Figure 2:

$$z'_t = z_t \odot (1 + \gamma_y) + \beta_y, \tag{2}$$

where γ_y and β_y denote the sentence-dependent scale and shift vectors, respectively, $\odot$ represents element-wise multiplication, and z'_t is the semantic-conditioned latent state. The modulated state z'_t, together with the time embedding t and E_{text}, is then fed into the Flow Matching Transformer for velocity prediction. SCLM and the flow predictor are trained with the standard conditional flow-matching objective, preserving the original interpolation path and velocity target. During inference, the trained controller generates γ_y and β_y from the input sentence, and the same transformation is applied to each current sampling state before velocity prediction. The residual design ensures that the original latent state is preserved when γ_y and β_y are zero.

3.2 Energy-Guided Trajectory Refinement

Flow-based samplers generate motions by iteratively following a learned velocity field, but the field does not explicitly measure whether intermediate states correspond to realistic sign motions. Errors in hand articulation, facial motion, joint configuration, or coordination among motion regions may therefore accumulate during sampling. To constrain these intermediate states, we propose Per-step Energy Guidance (PEG), an inference-time auxiliary guidance mechanism. As illustrated in the inference branch of Figure 2, PEG decodes a provisional latent state into motion space, evaluates its plausibility using a differentiable region-aware energy scorer, and feeds the resulting gradient

back to the latent state through the "Gradient Computation" and "Latent Correction" modules. PEG refines the sampling trajectory without replacing the velocity prediction of the flow generator.

3.2.1 Learning Region-aware Motion Energy

The energy scorer provides a differentiable motion-validity prior for PEG. As shown in Figure 3, a motion sequence is converted into kinematic features and partitioned into four components: facial motion, left-hand motion, right-hand motion, and inter-region interaction. Each component is processed by a regional MLP, and energies are combined by weighted aggregation. This factorized design enables the scorer to assess local motion quality and coordination among anatomical regions.

Specifically, let $\Psi(\cdot)$ denote a differentiable extractor of kinematic features, including joint positions, velocities, accelerations, and jerks. For a pose sequence X, the extracted features are divided into regional components

$$\Psi(X) = \{H_{\text{face}}, H_{\text{left}}, H_{\text{right}}, H_{\text{inter}}\},$$

where the first three terms describe facial, left-hand, and right-hand motion, respectively, and H_{inter} represents coordination features among different regions. For each region $r \in \mathcal{R}$, where $\mathcal{R} = \{\text{face}, \text{left}, \text{right}, \text{inter}\}$, a regional estimator f_r produces a scalar energy:

$$E_r(X) = f_r(H_r(X)), \qquad r \in \mathcal{R}. \tag{3}$$

The regional energies are then aggregated into the global motion energy

$$E_\omega(X) = \sum_{r \in \mathcal{R}} w_r E_r(X), \tag{4}$$

where w_r denotes the contribution weight of region r, and ω represents the parameters of the complete energy model.

Following the positive and negative branches in Figure 3, we construct a contrastive pair consisting of a realistic motion sequence X^+ and a corrupted sequence X^-. The positive sample is drawn from real sign-motion data, whereas the negative sample is generated through motion-specific perturbations, such as temporal disturbances, local articulation corruption, or joint swapping. The energy model is trained with a margin-based ranking loss:

$$\mathcal{L}_E = \max\left(0,\, m + E_\omega(X^+) - E_\omega(X^-)\right), \tag{5}$$

where $m > 0$ is a predefined margin. Minimizing this loss encourages realistic motions to receive lower energies than their corrupted counterparts. After training, the regional estimators and the aggregation function are frozen and used as the motion prior during inference.

3.2.2 Per-step Energy Guidance

During inference, PEG applies the energy prior to selected intermediate sampling states rather than only refining the final decoded motion. As shown in Figure 2, the provisional latent produced by the flow sampler is passed through the frozen VAE decoder. The decoded pose is then processed by the differentiable kinematic feature extractor and the Region-Aware Energy Scorer. The resulting energy gradient is computed with respect to the provisional latent and used by the "Latent Correction" module to obtain the refined latent for the next sampling step.

Let $z^{\text{prov}}_{t-\Delta t}$ denote a provisional state, P denote the mapping from the sampler state to the decoder latent space, and D_ϕ denote the pretrained and frozen VAE decoder. The motion-space energy of the provisional state is defined as

$$E_t = E_\omega\left(\Psi\left(D_\phi\left(P\left(z^{\text{prov}}_{t-\Delta t}\right)\right)\right)\right). \tag{6}$$

Although the energy is evaluated after decoding in motion space, the entire computation is differentiable. PEG therefore obtains a correction direction by

$$g_t = \nabla_{z^{\text{prov}}_{t-\Delta t}} E_t, \qquad \hat{g}_t = \frac{g_t}{\|g_t\|_{2,\text{feat}} + \epsilon}, \tag{7}$$

where ϵ is a small constant for numerical stability. The provisional state is then moved toward a lower-energy region:

$$z_{t-\Delta t} = z^{\text{prov}}_{t-\Delta t} - \lambda \hat{g}_t, \tag{8}$$

Table 1: Comparison with representative continuous sign language production methods on the PHOENIX-2014T test set. We report BLEU-1–4, ROUGE, CHRF, and DTW. Higher is better for BLEU, ROUGE, and CHRF, whereas lower is better for DTW. Bold and underlined values denote the best and second-best, respectively, among generated methods with available results. The GT row is an oracle reference and is excluded from the ranking. A dash denotes an unreported result.

Models	**ROUGE↑**	**BLEU-1↑**	**BLEU-2↑**	**BLEU-3↑**	**BLEU-4↑**	**CHRF↑**	**DTW↓**
GT	35.22	34.43	22.08	16.13	12.81	34.62	0.000
PT-base Saunders et al. (2020b)	14.52	6.27	3.33	2.14	1.59	9.50	0.197
PT-GN Saunders et al. (2020b)	19.09	11.45	7.08	5.08	4.04	14.52	0.191
G2P-MP Xie et al. (2024)	–	15.43	10.69	8.26	6.98	–	0.146
G2P-DDM Xie et al. (2024)	–	16.11	11.37	9.22	7.50	–	0.116
Sign-Base Walsh et al. (2025)	19.77	16.96	6.56	3.38	2.05	<u>25.88</u>	**0.040**
GCDM Tang et al. (2025b)	23.20	22.03	<u>14.21</u>	**10.16**	7.91	–	–
Sign-IDD Tang et al. (2025a)	<u>26.58</u>	<u>24.80</u>	–	–	**9.08**	–	–
Ours	**27.06**	**27.32**	**14.69**	<u>9.84</u>	<u>7.33</u>	**27.59**	<u>0.049</u>

Table 2: Component ablation study on the PHOENIX-2014T test set. SCLM denotes Semantic-Conditioned Latent Modulation, and PEG denotes Per-step Energy Guidance. Higher values are better for ROUGE, BLEU, and CHRF, whereas lower values are better for DTW.

Method	**ROUGE ↑**	**BLEU-1 ↑**	**BLEU-2 ↑**	**BLEU-3 ↑**	**BLEU-4 ↑**	**CHRF ↑**	**DTW ↓**
Baseline	21.36	22.36	10.05	5.88	4.06	22.87	0.058
Baseline + SCLM	22.58	23.61	11.57	7.31	5.18	23.71	0.057
Baseline + PEG	25.61	26.91	14.20	9.54	7.23	27.15	0.050
Baseline + SCLM + PEG	**27.06**	**27.32**	**14.69**	**9.84**	**7.33**	**27.59**	**0.049**

where λ controls the guidance strength. Thus, the flow sampler determines the primary generation direction, while PEG provides an auxiliary correction toward regions associated with realistic facial motion, hand articulation, and inter-region coordination.

PEG is activated only at selected sampling steps satisfying $t < \tau_{\mathrm{PEG}}$, and a predefined guidance ratio determines the frequency of correction among the eligible steps. At non-selected steps, the provisional state is passed to the next step without energy correction. After each correction, the updated state is detached before being used for subsequent sampling, preventing gradient accumulation through the complete trajectory. The energy scorer remains frozen during inference, and only its gradient with respect to the current provisional state is used by PEG.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate our method on the PHOENIX-2014T benchmark Camgoz et al. (2018), which contains 8,257 paired German spoken-language sentences and continuous 3D sign pose sequences. Following previous works Tang et al. (2025a); Shan et al. (2023); Xie et al. (2024), we use the training, development, and test splits, comprising 7,096, 519, and 642 paired samples, respectively. We additionally use the USTC-CSL dataset for cross-domain qualitative evaluation.

Evaluation Metrics. Following previous sign language production methods (Stoll et al. 2020; Shan et al. 2023; Tang et al. 2025a; Saunders, Camgoz, and Bowden 2020; Xie et al. 2024), we evaluate generated motions using recognition-based and motion-level metrics. Specifically, generated poses are translated back into text using a pretrained sign language translation model (Camgoz et al. 2018). We report BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE, and CHRF between the recognized and reference sentences. For motion-level evaluation, we use DTW-based Mean Joint Error (DTW) (Shan et al. 2023; Tevet et al. 2022). Higher values are preferred for BLEU, ROUGE, and CHRF, whereas lower values are preferred for DTW.

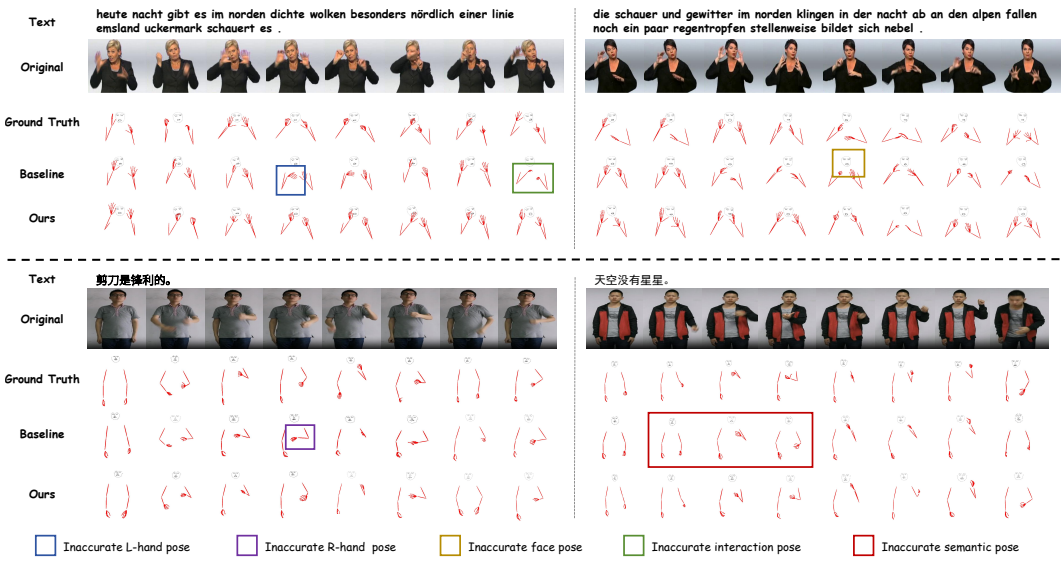

Figure 4: Qualitative comparisons on the PHOENIX-2014T (top) and USTC-CSL (bottom) datasets. Each example presents the text input, original video, ground-truth poses, baseline results, and our results from top to bottom, with colored boxes highlighting representative baseline errors in hand poses, facial motion, inter-region interaction, and text–motion semantics.

Implementation Details. Our framework follows a latent flow-based generation paradigm (Lipman et al. 2022; Rombach et al. 2022). A pretrained VAE encoder-decoder and T5 text encoder (Raffel et al. 2020) provide motion latents and textual features (Kingma and Welling 2013); both remain frozen during training. The *Baseline* is the latent conditional flow-matching generator without SCLM or PEG, using the same encoders and Transformer. We optimize the flow-matching backbone and SCLM with the conditional flow-matching objective. PEG is trained independently with a margin-based ranking loss on realistic and corrupted motion pairs and frozen during inference. At selected steps, it decodes the state into motion space, computes the energy from differentiable kinematic features, and applies the normalized energy gradient to correct the sampling state.

4.2 Comparison with Existing Methods

Table 1 compares SeMoSign with representative continuous sign language production methods on the PHOENIX-2014T test set. Among generated methods, SeMoSign achieves the best ROUGE, BLEU-1, BLEU-2, and CHRF scores, improving over the strongest prior results by approximately 1.8%, 10.2%, 3.4%, and 6.6%, respectively, indicating improved text–motion correspondence and semantic consistency. The GT row serves as an oracle reference and is excluded from comparisons. SeMoSign remains competitive on BLEU-3, BLEU-4, and DTW, ranking second among methods with DTW results, although GCDM, Sign-IDD, and Sign-base achieve the best BLEU-3, BLEU-4, and DTW scores, respectively. These varying rankings indicate that SeMoSign's advantages are not uniform across sequence-matching levels; moreover, lower DTW does not necessarily imply better semantic correspondence, as different motion trajectories may express the same linguistic content.

Figure 4 presents qualitative comparisons on the PHOENIX-2014T and USTC-CSL datasets. The regions reveal baseline errors in body-part poses and sequence-level semantics. On PHOENIX-2014T, the baseline produces inaccurate left-hand and interaction poses in the first example, marked by the blue and green boxes, and an inaccurate facial pose in the second, marked by the yellow box. On USTC-CSL, it shows an inaccurate right-hand pose in the first example and a sequence-level semantic mismatch in the second, indicated by the purple and red boxes. In contrast, SeMoSign produces more consistent hand, facial, and interaction poses while preserving semantics, demonstrating improved fine-grained pose modeling and text–motion correspondence.

4.3 Ablation Study

Table 3: Sensitivity analysis of the PEG guidance strength λ and guidance ratio ρ_{PEG} on the PHOENIX-2014T validation set. The selected values are fixed before evaluating the test set.

(a) Guidance strength λ

λ	ROUGE↑	BLEU-1↑	BLEU-4↑	DTW↓
0.01	25.97	26.98	6.61	0.049
0.10	**27.06**	**27.32**	**7.33**	**0.049**
0.50	22.40	23.06	4.85	0.056
1.00	21.82	22.72	4.34	0.057

Fixed ratio: $\rho_{\mathrm{PEG}} = 0.8$.

(b) Guidance ratio ρ_{PEG}

ρ_{PEG}	ROUGE↑	BLEU-1↑	BLEU-4↑	DTW↓
1.0	26.12	27.04	6.82	0.048
0.8	**27.06**	**27.32**	**7.33**	**0.049**
0.5	25.54	26.71	6.93	0.048
0.2	22.26	23.73	5.06	0.056

Fixed guidance strength: $\lambda = 0.1$.

Table 4: Regional ablation of PEG on the PHOENIX-2014T validation set; the full configuration is used for test evaluation.

Variant	ROUGE↑	BLEU-1↑	BLEU-4↑	DTW↓
Baseline+SCLM	22.58	23.61	5.18	0.057
w/o Face	25.75	26.56	6.99	0.048
w/o Left Hand	26.68	26.31	6.53	0.049
w/o Right Hand	26.40	26.69	6.71	0.049
w/o Interaction	25.35	26.16	6.11	0.049
Full PEG	**27.06**	**27.32**	**7.33**	**0.049**

Table 5: Inference efficiency. Latency is measured per sequence, throughput in sequences/s, and Δ relative to the base model; results are averaged over repeated runs under identical hardware and sampling settings.

Method	Lat.↓	Thr.↑	Δ(%)↓
B	1.3474 ± 0.4645	0.74	0.00
B+SCLM	1.3803 ± 0.4763	0.72	+2.45
B+PEG	1.4347 ± 0.6056	0.70	+6.48
B+SCLM+PEG	$\mathbf{1.4553 \pm 0.6280}$	**0.69**	**+8.01**

Component Ablation. Table 2 evaluates individual and combined effects of SCLM and PEG on the PHOENIX-2014T test set. PEG variants use hyperparameters and regional configuration selected on validation split. Both improve the baseline, with PEG gaining more. Combined, they achieve strongest results, supporting complementary roles: SCLM modifies sentence-conditioned input to the velocity predictor, whereas PEG corrects provisional states using a local motion prior.

Guidance Parameter Analysis. Table 3 analyzes the PEG guidance strength λ and guidance ratio ρ_{PEG} on the PHOENIX-2014T validation split. Moderate guidance achieves the best semantic performance, whereas overly strong guidance interferes with the flow-matching velocity field. A sparse guidance schedule weakens text–motion alignment. Based on the validation results, we set $\lambda = 0.1$ and $\rho_{\mathrm{PEG}} = 0.8$. These values are fixed before test-set evaluation and are used for all subsequent test-set experiments. The different trends between the semantic metrics and DTW indicate that the guidance schedule affects semantic fidelity and similarity to a reference trajectory differently.

Regional Energy Ablation. Table 4 evaluates the regional decomposition of the PEG scorer on the PHOENIX-2014T validation split. The scorer comprises facial, left-hand, right-hand, and cross-region interaction features. Each variant removes one component while keeping others. Removing any component weakens semantic performance, whereas the full configuration achieves the best recognition-based results, indicating that these feature groups provide complementary motion priors. Although DTW does not vary monotonically across variants, the full configuration is selected based on semantic performance on the validation split and fixed for subsequent test-set evaluations.

5 Conclusion

In this paper, we revisit flow-matching-based sign language production with energy guidance. Our framework jointly strengthens semantic representation learning during training and constrains the flow-based sampling trajectory during inference, enabling sign motions to remain semantically faithful and physically plausible. Extensive experiments on the PHOENIX-2014T benchmark demonstrate improvements in semantic- and motion-level metrics over strong flow-based baselines, while qualitative results verify the realism and naturalness of the generated motions. Since SeMoSign is largely orthogonal to existing flow-based SLP architectures, it offers potential for motion-prior-guided flow matching and inference-time trajectory refinement in human motion generation tasks.

AI USE STATEMENT

In this work, generative AI tools were used to assist with language polishing and LaTeX formatting. Generative AI tools were not used to generate experimental data, conduct experiments, produce reported results, or make unverifiable scientific claims. All AI-assisted text and formatting were reviewed and revised by the authors. The authors independently verified the technical content, implementation details, experimental results, and references, and take full responsibility for the final content of this paper, including any text or artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work utilizes two sign-language datasets released by other research teams. PHOENIX-2014T is a publicly available dataset, while USTC-CSL is a widely-used non-public official dataset within the research community. No new human-subject data was collected in this study. We follow the license terms for both datasets and do not release any personally identifiable information of participants. We acknowledge that existing sign-language datasets cannot fully represent the diversity of signers, signing styles, regional variations and linguistic backgrounds. Such limitations may affect the generalization and fairness of sign-language generation systems. The proposed method is intended solely for research on sign-language processing and cannot serve as a substitute for qualified human interpreters in high-stakes real-world scenarios.

REPRODUCIBILITY STATEMENT

The proposed method, training procedure, and inference pipeline are described in Section 3. Datasets, preprocessing procedures, evaluation metrics, and experimental protocols are detailed in Section 4. Additional implementation details, hyperparameter settings, ablation configurations, and supplementary qualitative visualizations are provided in Appendix A. All reported experimental results are obtained under the settings described within this paper. Full experimental settings, derivations and additional results are included in the paper appendix.

REFERENCES

Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. Neural sign actors: A diffusion model for 3d sign language production from text. *arXiv preprint arXiv:2312.02702*, 2023.

Richard Bowden. Deep sign: Hybrid cnn-hmm for continuous sign language recognition. In *Proceedings of the British Machine Vision Conference 2016*, 2016.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st international ACM SIGACCESS conference on computers and accessibility*, pp. 16–31, 2019.

Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, and Richard Bowden. Subunets: End-to-end hand shape and continuous sign language recognition. *ICCV 2017 Proceedings*, 2017.

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

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020*, 2020.

Yilun Du, Conor Durkan, Robin Strudel, Joshua B Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Sussman Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International conference on machine learning*, pp. 8489–8510. PMLR, 2023.

Biyi Fang, Jillian Co, and Mi Zhang. Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation. In *Proceedings of the 15th ACM conference on embedded network sensor systems*, pp. 1–13, 2017.

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

Jiashu He, Jiayi He, Shengeng Tang, Huixia Ben, Lechao Cheng, and Richang Hong. Stgcdiff: Spatial-temporal graph condition diffusion for sign language transition generation. In *Proceedings of the 3rd International Workshop on Deep Multimodal Generation and Retrieval*, pp. 11–20, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. Signbert: Pre-training of hand-model-aware representation for sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11087–11096, 2021.

Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li. Sign language recognition using 3d convolutional neural networks. In *2015 IEEE international conference on multimedia and expo (ICME)*, pp. 1–6. IEEE, 2015.

Eui Jun Hwang, Jung-Ho Kim, and Jong C Park. Non-autoregressive sign language production with gaussian space. In *BMVC*, volume 1, pp. 3, 2021.

Peiqi Jiao, Yuecong Min, and Xilin Chen. Visual alignment pre-training for sign language translation. In *European Conference on Computer Vision*, pp. 349–367. Springer, 2024.

Hamid Reza Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv preprint arXiv:1812.01053*, 2018.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2151–2162, 2023.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Oscar Koller, Sepehr Zargaran, and Hermann Ney. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4297–4305, 2017.

Oscar Koller, Necati Cihan Camgoz, Hermann Ney, and Richard Bowden. Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2306–2320, 2019.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The eleventh international conference on learning representations*, 2022.

John McDonald, Rosalee Wolfe, Jerry C Schnepp, Julie Hochgesang, Diana Gorman Jamrozik, Marie Stumbo, Larwan Berke, Melissa Bialek, and Farah Thomas. Automated technique for real-time production of lifelike animations of american sign language. *Universal Access in the Information Society*, 15(4), 2015.

Achraf Othman and Mohamed Jemni. Statistical sign language machine translation: from english written text to american sign language gloss. *arXiv preprint arXiv:1112.0168*, 2011.

Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European conference on computer vision*, pp. 480–497. Springer, 2022.

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

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

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Adversarial training for multi-channel sign language production. In *The 31st British Machine Vision Virtual Conference*. British Machine Vision Association, 2020a.

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

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

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

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Skeletal graph self-attention: Embedding a skeleton inductive bias into sign language production. In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pp. 95–102, 2022.

Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14761–14771, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. Text2sign: towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 128(4):891–908, 2020.

Shengeng Tang, Richang Hong, Dan Guo, and Meng Wang. Gloss semantic-enhanced network with online back-translation for sign language production. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 5630–5638, 2022a.

Shengeng Tang, Jiayi He, Dan Guo, Yanyan Wei, Feng Li, and Richang Hong. Sign-idd: Iconicity disentangled diffusion for sign language production. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7266–7274, 2025a.

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

Shiliang Tang, Dan Guo, Richang Hong, and Meng Wang. Graph-based multimodal sequential embedding for sign language translation. *IEEE Transactions on Multimedia*, pp. 4433–4445, 2022b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Carla Viegas, Mert Inan, Lorna Quandt, and Malihe Alikhani. Including facial expressions in contextual embeddings for sign language generation. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023)*, pp. 1–10, 2023.

Harry Walsh, Ed Fish, Ozge Mercanoglu Sincan, Mohamed Ilyes Lakhal, Richard Bowden, Neil Fox, Kearsy Cormier, 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, Meryem Tasyurek, Tugce Kiziltepe, and Hacer Yalim Keles. Slrtp2025 sign language production challenge: Methodology, results, and future work. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Pan Xie, Qipeng Zhang, Peng Taiying, Hao Tang, Yao Du, and Zexian Li. G2p-ddm: Generating sign pose sequence from gloss sequence with discrete diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6234–6242, 2024.

Kayo Yin. Sign language translation with transformers. *arXiv preprint arXiv:2004.00588*, 2, 2020.

Jan Zelinka and Jakub Kanis. Neural sign language synthesis: Words are our glosses. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3395–3403, 2020.

Jianrong Zhang, Hehe Fan, and Yi Yang. Demogen: Towards decompositional human motion generation with energy-based diffusion models. *arXiv preprint arXiv:2512.22324*, 2025a.

Jianrong Zhang, Hehe Fan, and Yi Yang. Energymogen: Compositional human motion generation with energy-based diffusion model in latent space. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17592–17602, 2025b.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4115–4128, 2024.

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20871–20881, 2023.

A Appendix

A.1 Additional Technical Details

This section provides the additional formulations omitted from the main paper for compactness. We focus on the construction of the kinematic features, the regional energy scorer, and the implementation of Per-step Energy Guidance (PEG). The overall flow-matching and semantic-conditioning formulations remain the same as those described in the main paper.

A.1.1 Kinematic Feature Construction

Let

$$X = \{X_t\}_{t=1}^{T}, \qquad X_t \in \mathbb{R}^{J\times d},$$

denote a pose sequence with J joints and d-dimensional joint coordinates. In addition to joint positions, PEG uses temporal kinematic features to characterize local motion dynamics. We compute velocity, acceleration, and jerk using finite differences:

$$V_t = \Delta X_t = X_t - X_{t-1}, \tag{9}$$

$$A_t = \Delta V_t = V_t - V_{t-1}, \tag{10}$$

$$K_t = \Delta A_t = A_t - A_{t-1}. \tag{11}$$

The features are temporally aligned with the original pose sequence. We denote the complete kinematic representation by

$$\Psi(X) = \{X, V, A, K\}.$$

This representation allows the energy model to assess not only instantaneous joint configurations but also motion smoothness and higher-order temporal variations.

A.1.2 REGIONAL DECOMPOSITION

Let $\mathcal{J}_{\text{face}}$, $\mathcal{J}_{\text{left}}$, and $\mathcal{J}_{\text{right}}$ denote the facial, left-hand, and right-hand joint sets, respectively. For a region r, the corresponding regional feature is obtained by selecting the associated joints from the global kinematic representation:

$$H^r(X) = \text{Select}_{\mathcal{J}_r}\big(\Psi(X)\big), \qquad r \in \{\text{face}, \text{left}, \text{right}\}.$$

The fourth component describes coordination between anatomical regions. In general, given a set of cross-region joint pairs

$$\mathcal{P}_{\text{inter}} \subseteq \mathcal{J} \times \mathcal{J},$$

we construct relative position and velocity features as

$$R^p_{t,(i,j)} = X_{t,i} - X_{t,j}, \tag{12}$$

$$R^v_{t,(i,j)} = V_{t,i} - V_{t,j}, \qquad (i,j) \in \mathcal{P}_{\text{inter}}. \tag{13}$$

The interaction feature is then written as

$$H^{\text{inter}}(X) = \left[R^p, R^v\right].$$

Therefore, the four feature groups used by PEG are

$$\Psi(X) \longrightarrow \left\{H^{\text{face}}(X), H^{\text{left}}(X), H^{\text{right}}(X), H^{\text{inter}}(X)\right\}.$$

This decomposition separates local articulation from coordination patterns between different motion regions.

A.1.3 REGIONAL ENERGY SCORING

Each regional feature is processed by an independent energy estimator:

$$E_r(X) = f_r(H^r(X); \theta_r), \qquad r \in \mathcal{R},$$

where

$$\mathcal{R} = \{\text{face}, \text{left}, \text{right}, \text{inter}\}.$$

The global motion energy is obtained by weighted aggregation:

$$E_\omega(X) = \sum_{r \in \mathcal{R}} w_r E_r(X), \tag{14}$$

where w_r controls the contribution of each motion region and ω denotes the complete set of scorer parameters. For a realistic motion X^+ and a corrupted motion X^-, the scorer is trained with the margin-ranking objective

$$\mathcal{L}_E = \max\left(0,\, m + E_\omega(X^+) - E_\omega(X^-)\right), \tag{15}$$

where $m > 0$ is the margin. Minimizing $\mathcal{L}_E$ encourages realistic motions to occupy lower-energy regions than their corrupted counterparts. The corrupted motion is generated by applying a motion perturbation operator to a real sequence:

$$X^- = \mathcal{C}_q(X^+), \qquad q \in \mathcal{Q}, \tag{16}$$

where $\mathcal{Q}$ denotes the set of corruption types, such as temporal disturbances, local articulation perturbations, and joint-level rearrangements. The energy scorer is trained separately from the flow-matching generator and is frozen after optimization.

A.1.4 DIFFERENTIABLE PEG CORRECTION

At a sampling step, let $z^{\text{prov}}_{t-\Delta t}$ denote the provisional latent produced by the flow-matching update. The corresponding motion-space pose is

$$X^{\text{prov}}_{t-\Delta t} = D_\phi\left(P\left(z^{\text{prov}}_{t-\Delta t}\right)\right), \tag{17}$$

where P maps the sampler state to the decoder input space and D_ϕ is the frozen VAE decoder. PEG evaluates

$$E_t = E_\omega\left(X^{\text{prov}}_{t-\Delta t}\right). \tag{18}$$

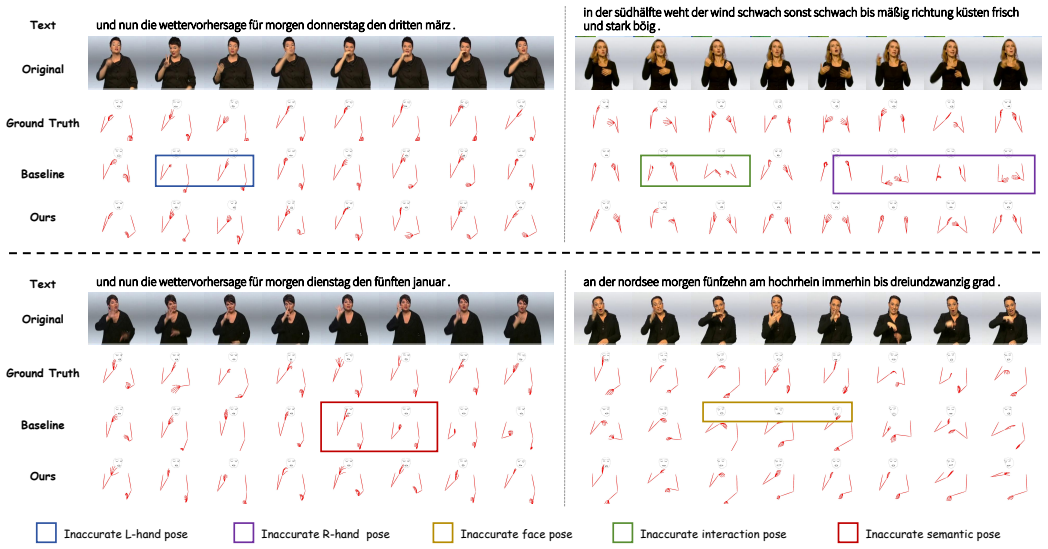

Figure 5: Additional qualitative comparisons on PHOENIX-2014T. Each example shows the input text, ground-truth motion, baseline output, and SeMoSign output.

Because the decoder, kinematic feature extractor, and regional energy scorer are differentiable, the correction gradient can be computed by the chain rule:

$$g_t = \frac{\partial E_\omega}{\partial X}\frac{\partial X}{\partial z}\bigg|_{z=z_{t-\Delta t}^{\text{prov}}} = \frac{\partial E_\omega}{\partial X}\frac{\partial D_\phi(P(z))}{\partial z}\bigg|_{z=z_{t-\Delta t}^{\text{prov}}} \tag{19}$$

We normalize the gradient along the latent feature dimension:

$$\widehat{g}_t = \frac{g_t}{\|g_t\|_{2,\text{feat}} + \epsilon}, \tag{20}$$

and update the provisional state toward a lower-energy region:

$$z_{t-\Delta t} = z_{t-\Delta t}^{\text{prov}} - \lambda \widehat{g}_t. \tag{21}$$

PEG is activated only at eligible steps satisfying $t < \tau_{\text{PEG}}$. Let ρ_{PEG} denote the guidance ratio and let $\mathcal{S}_{\text{PEG}}$ be the selected set of correction steps:

$$\mathcal{S}_{\text{PEG}} \subseteq \{t : t < \tau_{\text{PEG}}\}, \qquad \frac{|\mathcal{S}_{\text{PEG}}|}{|\{t : t < \tau_{\text{PEG}}\}|} = \rho_{\text{PEG}}. \tag{22}$$

The complete update can therefore be written as

$$z_{t-\Delta t} = \begin{cases} z_{t-\Delta t}^{\text{prov}} - \lambda \widehat{g}_t, & t \in \mathcal{S}_{\text{PEG}}, \\ z_{t-\Delta t}^{\text{prov}}, & t \notin \mathcal{S}_{\text{PEG}}. \end{cases} \tag{23}$$

After each correction, the updated state is detached before being passed to the next sampling step. Consequently, PEG modifies the sampling trajectory locally without back-propagating gradients through the complete generation process. In our experiments, we use $\lambda = 0.1$ and $\rho_{\text{PEG}} = 0.8$.

A.2 Additional Qualitative Results

This appendix provides additional qualitative examples on PHOENIX-2014T. The examples complement the comparisons in the main paper by illustrating hand articulation, facial motion, inter-region coordination, and sequence-level semantics.

A.2.1 Additional Visualizations

Figure 5 shows supplementary comparisons on PHOENIX-2014T. The examples cover hand articulation, facial motion, inter-region coordination, and sequence-level semantic consistency.