Spectral Optimal Transport for Dynamics-Aware Sign Language Generation

Nabeela Khan[1,2], Bowen Wu[2,3], Runwu Shi[1], Hanlong Li[1], Benjamin Yen[4,1], Carlos Toshinori Ishi[2,3], and Kazuhiro Nakadai[1]

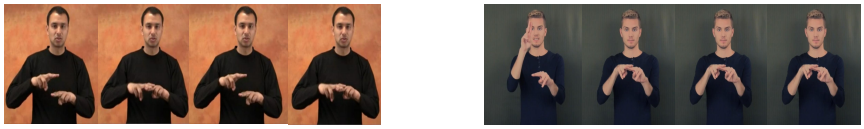
[1] Institute of Science Tokyo, Tokyo, Japan
{nabeela,shirunwu,lihanlong,nakadai}@ra.sc.eng.isct.ac.jp
[2] RIKEN Guardian Robot Project, Kyoto, Japan
{bowen.wu,carlos.ishi}@riken.jp
[3] ATR Hiroshi Ishiguro Laboratories, Kyoto, Japan
[4] RIKEN Center for Biosystems Dynamics Research, Kobe, Japan
benjamin.yen@ieee.org

Abstract. Sign language generation (SLG) translates spoken-language text into pose sequences that convey its meaning. Recently, flow-based models with optimal transport (OT) have become a strong SLG baseline by handling the one-to-many nature of text-to-sign. While classifier-free guidance (CFG) could be applied to further strengthen the semantic alignment theoretically, its training involves conditional and unconditional training, which may contradict the pose-space OT, thus reducing the generation quality. Under the condition dropout, one noise sample can be paired with target motions exhibiting incompatible motion dynamics, interfering with the semantic guidance when the text condition is present. To address this issue, we propose Spectral Optimal Transport (SOT), which augments the pose-space cost with a spectral feature distance, reweighting the coupling toward the temporal-frequency structure of the target motions, thereby improving the performance when using CFG. The quantitative and qualitative experiments demonstrate the effectiveness of the proposed method. The code is available at https://github.com/foxrei-n/SignFlow.

Keywords: Sign language generation · Conditional flow matching · Classifier-free guidance · Motion dynamics modeling

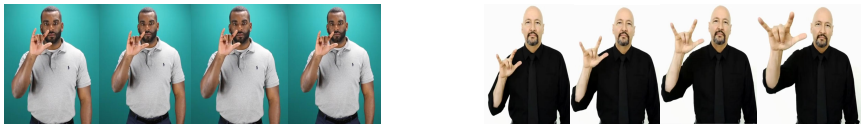
1 Introduction

Sign languages (SLs) are full-fledged natural languages used by Deaf and Hard-of-Hearing (DHH) communities [50]. They are expressed in the visual–manual modality through coordinated manual and non-manual elements [33]. Sign language generation (SLG) aims to generate sign pose sequences conditioned on spoken-language text, where a central goal is *semantic alignment*, meaning the generated motion should convey the intended meaning. Prior work has improved SLG by enhancing motion realism [3], reducing prediction drift in long sequences [17], handling information density [49], and incorporating expressive



"Chair" *"Sit"*

(a) Repetition vs. single motion



"Airplane" *"Fly"*

(b) Small vs. large path extent

Fig. 1: Minimal pairs illustrating the phonological role of motion dynamics in sign language. The images are sourced from WLASL [21].

factors such as emotions [29]. Recently, SignFlow [18] emerged as a strong flow-based SLG model that better handles the *one-to-many* nature of SLG via conditional flow matching and minibatch optimal transport (CFM-OT). However, because the same text can correspond to multiple valid realizations, the text alone does not uniquely determine the pose sequence at inference. To improve adherence to the input text, we use the standard classifier-free guidance (CFG) procedure [13]. Empirically, we confirm that CFG improves semantic alignment; however, improved semantic alignment does not necessarily guarantee that the generated motion realizes the motion dynamics required for linguistically meaningful signing. Generally, we view such motion dynamics as part of aligned semantic realization. SLs exhibit a structured phonology [8, 37] where signs are composed of sublexical features such as handshape, location, movement, and orientation [8], and meaning can depend critically on *how motion evolves over time* (e.g., repetition vs. single movement, fast vs. slow articulation) [5]. Figure 1 illustrates minimal contrasts where the main difference is motion dynamics: *chair* vs. *sit*, differs by repeated vs. single movement, and *airplane* vs. *fly* differs by repeated small motions vs. a larger single motion. These cues are not merely stylistic; suppressing them can yield visually plausible but linguistically incorrect signs [15]. Importantly for SLG, such dynamics also vary across signers and contexts, so even identical text can correspond to multiple valid realizations with distinct motion dynamics.

In SignFlow, minibatch OT computes a permutation that pairs prior noise samples with target motion sequences, shaping the supervision for the learned vector field [18]. CFM-OT [44] uses a Euclidean pose-space coupling cost that is largely insensitive to motion dynamics. This becomes problematic under CFG-style condition dropout, where the same network is trained on both unmasked (text-conditioned) and masked (unconditioned) examples. Conceptually, unmasked updates are driven by a conditional motion distribution that is relatively concentrated for a given sentence, while masked updates are driven by a broader

marginal motion distribution that mixes diverse dynamics across the dataset. This mismatch makes the masked and unmasked training signals less compatible, which can limit the effectiveness of guidance at inference when the two branches are combined.

We therefore propose Spectral Optimal Transport (SOT), a frequency-aware OT coupling that augments the assignment cost with a spectral distance computed from temporal DCT (Discrete Cosine Transform) features, making the assignment sensitive to the temporal-frequency structure of the target motions. Prior work has modeled motion dynamics (trajectories) with DCT bases [1], and later motion prediction methods used frequency-space formulations, often to promote smoothness [26, 43]. In SLG, however, preserving motion dynamics (e.g., repetition and path extent) is important for faithful realizations, and these dynamics may vary across valid sequences for the same text. Rather than comparing motion dynamics between the two operands, a prior sample has none; SOT reweights the metric under which noise and targets are coupled: mismatches in higher temporal-frequency components contribute more to the assignment cost, so targets that differ in motion dynamics become more distinguishable during coupling. This reduces masked–unmasked inconsistency under CFG training and amplifies the gains from guidance.

This paper makes the following contributions:

- We empirically demonstrated that classifier-free guidance (CFG) improves semantic alignment for flow-based text-to-sign language generation.
- We proposed Spectral Optimal Transport (SOT), a simple frequency-aware extension of minibatch OT that reweights the coupling cost toward temporal-frequency structure, so that the assignment becomes sensitive to the motion dynamics of the targets.
- We showed that SOT reduces masked-unmasked mismatch and amplifies the gains from CFG.

2 Related Work

2.1 Sign Language Generation

A large number of SLG works rely on *glosses*[5] as an intermediate representation, following a text→gloss→pose (or video) pipeline [14,36,40,41,45–47,51,55]. While glosses can simplify learning by providing a discrete gloss-token interface, they require costly expert annotation and introduce an intermediate bottleneck that may discard linguistically relevant variation beyond the label sequence [27]. These constraints have motivated gloss-free approaches that map spoken-language text more directly to sign pose sequences. Early work in this direction explored transformer-based continuous sign language production [17, 34, 35]. Beyond spoken-language text, Ham2Pose maps HamNoSys sign notation to pose

[5] A gloss is a label representing an isolated sign, often used as an intermediate symbolic representation.

sequences, demonstrating that structured linguistic inputs can be animated directly into continuous signing motion [2]. More recent works tried to improve sentence-level semantic faithfulness via learned motion representations and decoding strategies, including discrete pose codebooks with vector quantization [15], non-autoregressive generation with iterative latent refinement [19], diffusion-based generation in learned latent spaces [3,10,11], and flow-matching-based generation [18]. Overall, these works reduce dependency on gloss annotations and improve sentence-level semantic alignment. However, most methods focus on matching joint configurations and do not explicitly model temporal dynamics, such as repetition or path extent, which can hinder the capture of the diversity of valid realizations in SLG.

2.2 Guidance and Coupling in Conditional Generation

Classifier-free guidance (CFG) is widely used to strengthen conditioning in generative models. Popularized in diffusion models, CFG combines unconditional and conditional predictions during sampling with a guidance scale that increases adherence to the conditioning signal while typically reducing diversity at high scales [13]. It has since been adopted broadly in conditional generation, including text-to-image and text-to-motion synthesis, because it often improves semantic alignment without modifying the underlying model architecture [30–32,42]. More recently, guidance mechanisms have also been studied for flow-based generative models. Zheng et al. [52] introduced *guided flows*, framing guidance as an explicit mechanism to steer continuous-time dynamics for both generation and downstream decision-making. Fan et al. [9] revisited CFG in the flow-matching setting and proposed variants tailored to flow models, highlighting that guidance behavior in flows can differ from diffusion. Complementarily, Feng et al. [12] provided a systematic treatment of guidance in flow matching, unifying multiple guidance formulations and clarifying their connections and trade-offs. Despite its effectiveness, CFG primarily strengthens conditional adherence at inference, and its success depends on learning compatible unconditional and conditional fields under condition dropout. In OT-trained flow matching, the training signal is shaped by minibatch couplings; recent work has shown that naive couplings in conditional settings can induce train–test mismatch when the coupling cost ignores structure relevant to the condition [7]. Our work connects these threads by studying CFG in CFM-OT for SLG and introducing a coupling strategy that incorporates temporal spectral structure to improve the compatibility of conditional and unconditional training signals.

Motivation: Dynamics-aware couplings under CFG training. CFG training alternates between unconditional (masked) and conditional (unmasked) updates, which emphasize different motion distributions ($p(x)$ vs. $p(x \mid c)$) [13]. With pose-only OT costs, minibatch couplings can mix samples with incompatible dynamics, especially in masked updates, leading to inconsistent pairings across the two training modes. We therefore propose **Spectral Optimal Transport (SOT)**, which augments the OT cost with frequency-domain motion features so

Algorithm 1 Minibatch OT Coupling (Euclidean)

Require: Source distribution p_0, targets $\mathbf{X}_1 = \{\mathbf{x}_1^j\}_{j=1}^B$, conditions $\mathbf{Y} = \{\mathbf{y}^j\}_{j=1}^B$
Ensure: Coupled pairs $(\mathbf{X}_0, \mathbf{X}_1', \mathbf{Y}')$
Sample $\mathbf{X}_0 = \{\mathbf{x}_0^i\}_{i=1}^B \sim p_0$
Compute $M_{\text{pose}}(i,j) = \|\text{vec}(\mathbf{x}_0^i) - \text{vec}(\mathbf{x}_1^j)\|_2^2$
Solve assignment $\pi \leftarrow \arg\min_\pi \sum_{i=1}^B M_{\text{pose}}(i, \pi(i))$ ▷ Hungarian
Permute: $\mathbf{X}_1' = \{\mathbf{x}_1^{\pi(i)}\}_{i=1}^B$, $\mathbf{Y}' = \{\mathbf{y}^{\pi(i)}\}_{i=1}^B$
return $\mathbf{X}_0, \mathbf{X}_1', \mathbf{Y}'$

Algorithm 2 Spectral OT Coupling (SOT, Ours)

Require: Targets $\mathbf{X}_1 = \{\mathbf{x}_1^j\}_{j=1}^B$, conditions $\mathbf{Y} = \{\mathbf{y}^j\}_{j=1}^B$
Require: Source distribution p_0, spectral ratio ρ, descriptor $\psi(\cdot)$
Ensure: Coupled pairs $(\mathbf{X}_0, \mathbf{X}_1', \mathbf{Y}')$
Sample $\mathbf{X}_0 = \{\mathbf{x}_0^i\}_{i=1}^B \sim p_0$
Compute $M_{\text{pose}}(i,j) = \|\text{vec}(\mathbf{x}_0^i) - \text{vec}(\mathbf{x}_1^j)\|_2^2$
Compute $M_{\text{spec}}(i,j) = \|\psi(\mathbf{x}_0^i) - \psi(\mathbf{x}_1^j)\|_2^2$
Set $\lambda_{\text{spec}} \leftarrow \rho \cdot \dfrac{\mathbb{E}[M_{\text{pose}}]}{\mathbb{E}[M_{\text{spec}}] + \epsilon}$
Combine $M(i,j) \leftarrow M_{\text{pose}}(i,j) + \lambda_{\text{spec}} M_{\text{spec}}(i,j)$
Solve assignment $\pi \leftarrow \arg\min_\pi \sum_{i=1}^B M(i, \pi(i))$ ▷ Hungarian
Permute: $\mathbf{X}_1' = \{\mathbf{x}_1^{\pi(i)}\}_{i=1}^B$, $\mathbf{Y}' = \{\mathbf{y}^{\pi(i)}\}_{i=1}^B$
return $\mathbf{X}_0, \mathbf{X}_1', \mathbf{Y}'$

that couplings respect both pose similarity and temporal dynamics, improving CFG-based generation.

3 Method

3.1 Preliminaries

Conditional flow matching. Flow Matching (FM) and its conditional variant (CFM) learn a time-dependent vector field by regressing to a target velocity along a chosen probability path [24]. Let $\mathbf{x}_1 \in \mathbb{R}^d$ denote a target SL motion sequence (e.g., a T-frame pose sequence flattened into a d-dimensional vector), and let $\mathbf{y}$ be the text condition. We denote by $\mathbf{x}(t) \in \mathbb{R}^d$ an intermediate point along a chosen probability path that transports samples from a simple source distribution to the data distribution, and by $\mathbf{v}_\theta(t, \mathbf{x} \mid \mathbf{y}) \in \mathbb{R}^d$ a neural, time-dependent conditional velocity field parameterized by θ. We draw an initial sample $\mathbf{x}_0 \sim p_0$ from a simple source distribution ($p_0 = \mathcal{N}(\mathbf{0}, \mathbf{I})$), and generate samples by solving the conditional ODE

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{v}_\theta\big(t, \mathbf{x}(t) \mid \mathbf{y}\big), \qquad \mathbf{x}(0) = \mathbf{x}_0. \tag{1}$$

Minibatch optimal transport (OT) coupling. To obtain straight probability paths, a minibatch of noise samples is coupled with a minibatch of data samples using OT, replacing random pairing with a cost-minimizing permutation [44]. Given noise $\mathbf{X}_0 = \{\mathbf{x}_0^i\}_{i=1}^B$ and data $\mathbf{X}_1 = \{\mathbf{x}_1^j\}_{j=1}^B$, OT computes a permutation $\mathbf{P} \in \mathcal{P}_B$ that minimizes a pairwise cost, typically Euclidean distance in pose space. After coupling, each paired $(\mathbf{x}_0, \mathbf{x}_1)$ defines the linear path

$$\mathbf{x}_t \;=\; \phi_t(\mathbf{x}_0, \mathbf{x}_1) \;=\; \big(1 - (1 - \sigma_{\min})t\big)\mathbf{x}_0 + t\,\mathbf{x}_1, \tag{2}$$

with the corresponding ideal (time-invariant) velocity

$$\mathbf{u}_t^{\mathrm{OT}}(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_1) \;=\; \mathbf{x}_1 - (1 - \sigma_{\min})\mathbf{x}_0. \tag{3}$$

CFM then regresses $\mathbf{v}_\theta\big(t, \mathbf{x}_t \mid \mathbf{y}\big)$ to $\mathbf{u}_t^{\mathrm{OT}}$ along these OT-coupled paths.

Classifier-free guidance (CFG) for flow matching. We adopt a CFG [13] condition dropout scheme to train a single model that supports both conditional and unconditional vector fields. Let $b \sim \mathrm{Bernoulli}(p_{\mathrm{uncond}})$ and define the dropped condition

$$\tilde{\mathbf{y}} = \begin{cases} \mathbf{y}, & b = 0, \\ \emptyset, & b = 1, \end{cases} \tag{4}$$

where $\emptyset$ denotes the null condition. Under the OT-coupled linear path $\mathbf{x}_t = \phi_t(\mathbf{x}_0, \mathbf{x}_1)$, we optimize

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{t,\, b,\, (\mathbf{x}_1, \mathbf{y}) \sim q,\, \mathbf{x}_0 \sim p_0}\Big[\big\|\mathbf{v}_{\boldsymbol{\theta}}\big(t, \mathbf{x}_t \mid \tilde{\mathbf{y}}\big) - \mathbf{u}_t^{\mathrm{OT}}\big(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_1\big)\big\|_2^2\Big]. \tag{5}$$

At inference time, guidance combines unconditional and conditional predictions as

$$\tilde{\mathbf{v}}_{\boldsymbol{\theta}}(t, \mathbf{x} \mid \mathbf{y}) = \mathbf{v}_{\boldsymbol{\theta}}(t, \mathbf{x} \mid \emptyset) + \omega\Big(\mathbf{v}_{\boldsymbol{\theta}}(t, \mathbf{x} \mid \mathbf{y}) - \mathbf{v}_{\boldsymbol{\theta}}(t, \mathbf{x} \mid \emptyset)\Big). \tag{6}$$

where $\omega \geq 0$ is the guidance scale (typically $\omega > 1$ for stronger conditioning), and we integrate the ODE from $t = 0$ to $t = 1$ using $\tilde{\mathbf{v}}_{\boldsymbol{\theta}}$.

3.2 Spectral Optimal Transport Coupling

We introduce **Spectral Optimal Transport (SOT)**, a frequency-aware minibatch OT coupling used during CFM-OT training. SOT modifies the minibatch assignment cost by combining a standard pose-space distance with a temporal spectral distance, so that the coupling accounts for the targets' temporal-frequency structure in addition to pose similarity.

Procedure. Alg. 2 summarizes SOT. For each minibatch $(\mathbf{X}_1, \mathbf{Y}) = \{(\mathbf{x}_1^j, \mathbf{y}^j)\}_{j=1}^B \sim q(\mathbf{x}_1, \mathbf{y})$, we (i) sample prior noise $\mathbf{X}_0 = \{\mathbf{x}_0^i\}_{i=1}^B \sim p_0$, (ii) compute a pose-space cost matrix M_{pose}, (iii) compute a temporal spectral descriptor $\psi(\cdot)$ for both $\mathbf{X}_0$ and $\mathbf{X}_1$ to obtain a spectral cost matrix M_{spec}, (iv) scale the spectral term via ratio normalization, and (v) solve a minimum-cost assignment with Hungarian matching [20] to obtain the coupling permutation. The resulting paired samples are then used to construct the CFM-OT training paths and target velocities (Sec. 3.1).

Temporal DCT descriptor. We instantiate $\psi(\mathbf{x})$ with a temporal DCT because it provides a compact, global summary of periodicity and smoothness in motion trajectories, is efficient and hyperparameter-free, and has been widely used in human motion modeling [1, 22, 25, 26, 48]. Given a pose sequence $\mathbf{x} \in \mathbb{R}^{T\times D}$, we compute a DCT-II along time for each feature dimension d:

$$\Psi_{k,d}(\mathbf{x}) = \sqrt{w_k} \sum_{t=0}^{T-1} x_{t,d} \cos\Big(\frac{\pi}{T}\big(t+\tfrac{1}{2}\big)k\Big), \quad k = 0, \dots, T-1, \tag{7}$$

where $w_k = 1+(\beta-1)\,k/(T-1)$ is a frequency-dependent gain increasing linearly from 1 at $k=0$ to β at $k=T-1$ ($\beta=3$ in our experiments), and we form $\psi(\mathbf{x})$ by concatenating coefficients $\Psi_{k,d}(\mathbf{x})$ across k and d. The gain places more weight on faster-varying temporal components, which would otherwise contribute little to the coupling cost since motion energy concentrates at low frequencies.

Spectral OT cost. Using $\psi(\cdot)$, SOT defines the pose and spectral costs

$$M_{\text{pose}}(i,j) = \Big\|\text{vec}(\mathbf{x}_0^i) - \text{vec}(\mathbf{x}_1^j)\Big\|_2^2, \tag{8}$$

$$M_{\text{spec}}(i,j) = \Big\|\psi(\mathbf{x}_0^i) - \psi(\mathbf{x}_1^j)\Big\|_2^2, \tag{9}$$

and couples minibatches using the combined cost

$$M(i,j) = M_{\text{pose}}(i,j) + \lambda_{\text{spec}}\, M_{\text{spec}}(i,j). \tag{10}$$

What the spectral term does to the coupling. Instead of comparing the motion dynamics of noise, which is a prior sample, with those of a target, SOT reweights the metric under which the two are coupled: since a minimum-cost assignment is invariant to row- and column-constant terms of the quadratic cost in Eq. (10), the coupling is determined entirely by the frequency-weighted correlation between the DCT coefficients of $\mathbf{x}_0$ and $\mathbf{x}_1$, with the weight of frequency k set by the gain w_k. Under a pose-only cost, all temporal frequencies are weighted equally, so, because sign-motion energy concentrates at low frequencies, the assignment is decided almost entirely by the slow, gross trajectory. The gain w_k raises the share contributed by the components that carry repetition and articulation speed, making targets that differ in motion dynamics more distinguishable during coupling. The motion dynamics thus enter on the *target* side throughout: since $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is isotropic and ψ is linear, $\psi(\mathbf{x}_0)$ acts as a dynamics-neutral probe. In practice, the spectral term is a small perturbation of the pose cost that mainly biases how near-tied assignments are resolved toward pairings consistent with the targets' temporal-frequency structure, which also yields more stable minibatch couplings.

Ratio-normalized spectral weighting. Because M_{pose} and M_{spec} can differ in scale, we set

$$\lambda_{\text{spec}} = \rho \cdot \frac{\mathbb{E}[M_{\text{pose}}]}{\mathbb{E}[M_{\text{spec}}] + \epsilon}, \tag{11}$$

Table 1: Quantitative results DTW-based pose metrics (PA-JPE and JPE) [54]. BLEU-4 (B-4) represents a semantic metric. Since pretrained checkpoints are not publicly released for the evaluator and some baselines, we train the required models from scratch.

Method	CSL-Daily					Phoenix-2014T				
	PA-JPE↓		JPE↓		B-T↑	PA-JPE↓		JPE↓		B-T↑
	Body	Hand	Body	Hand	B-4	Body	Hand	Body	Hand	B-4
Prog. Trans. [34]	15.98	12.91	16.30	32.63	1.97	13.67	11.95	15.01	31.77	2.03
Text2Mesh [38]	13.47	12.10	13.76	30.37	–	13.48	12.06	14.04	31.64	–
T2S-GPT [49]	11.94	5.93	12.32	15.43	–	10.38	6.47	11.65	19.09	–
S-MotionGPT [16]	10.81	3.78	11.58	11.31	–	9.45	3.41	10.42	9.08	–
SOKE [54]	6.24	1.71	7.38	9.68	6.22	4.77	1.38	6.04	7.72	6.92
SignFlow [18]	4.49	1.03	7.45	4.51	6.84	4.47	1.00	4.78	6.52	8.91
Ours	**4.42**	**1.01**	**6.87**	**4.43**	**9.09**	**4.32**	**0.97**	**4.76**	**6.06**	**9.78**

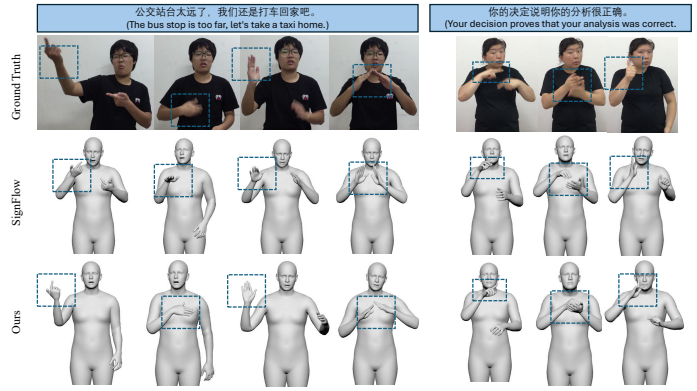
where $\mathbb{E}[\cdot]$ denotes the minibatch mean, ρ controls the target relative contribution of the spectral term, and ϵ is a small constant for numerical stability.

Coupling and usage. We solve the minimum-cost assignment under M (Eq. 10) using Hungarian matching (Alg. 2) to obtain a coupling permutation. The paired $(\mathbf{x}_0, \mathbf{x}_1)$ are then used to construct the CFM-OT training path and target velocity (Sec. 3.1).

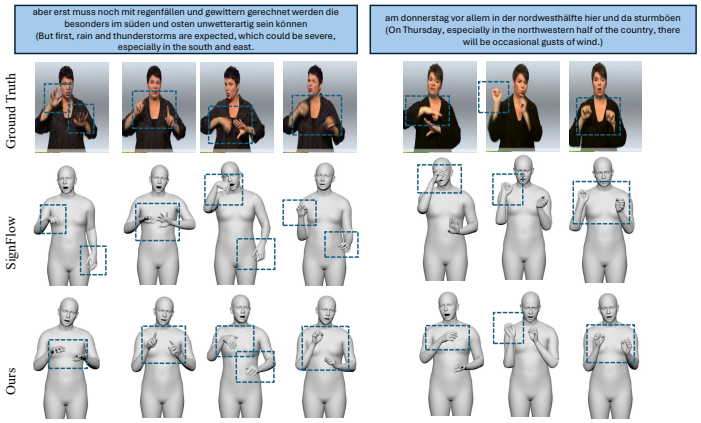
4 Experiments

We conduct experiments on two widely used SLG benchmark datasets: CSL-Daily [53] and Phoenix-2014T [6]. CSL-Daily is a large-scale Chinese Sign Language dataset of daily conversations with paired spoken-language sentences and sign videos (about 20K samples). Phoenix-2014T is a German Sign Language benchmark of weather forecast news (about 8K samples). For all experiments, we represent signing as SMPL-X pose sequences. We use SMPL-X motion features from the curated preprocessing released by [54], using 6D rotations and 10 SMPL-X shape parameters.

Following [54], we evaluate generated sign pose sequences by first aligning each generated sequence to its ground-truth counterpart using Dynamic Time Warping (DTW) [4]. We report DTW-aligned joint position errors under both Procrustes alignment (PA-JPE) and the original joint coordinates (JPE), which quantify sequence-level distances between generated motions and ground truth [2]. Following prior SLG work [3], we evaluate semantic alignment using Back-Translation (BT): a separate sign-to-text model translates each generated pose sequence back into spoken-language text, and we compute ROUGE-L [23]



(a) CSL-Daily examples.

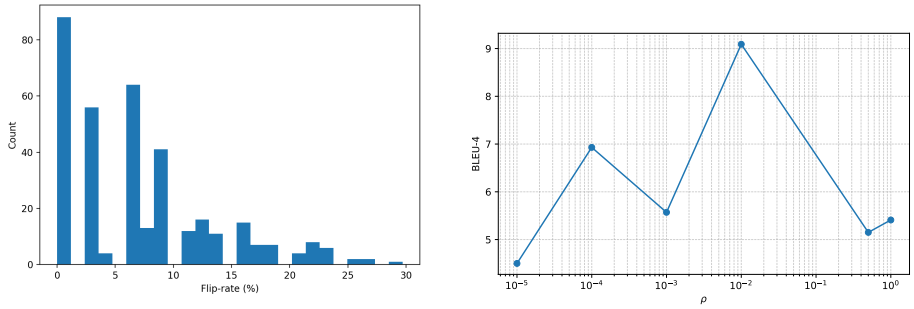


(b) Phoenix-2014T examples.

Fig. 2: Qualitative comparison. We show representative examples from CSL-Daily and Phoenix-2014T.

and BLEU-n [28] between the reconstructed and reference texts. We train the sign-to-text model of [39] as our BT evaluator; training details are provided in the supplementary material.

Implementation details. We build on the CFM-OT baseline [18]. Across all experiments, we keep the architecture, text encoder, training schedule, and optimization settings fixed as mentioned in [18]. For OT coupling, we assume uniform marginals and compute the minimum-cost bipartite assignment using the Hungarian algorithm [20]. For SOT, we set the pose-space term weight to 1 and compute the temporal DCT descriptor by controlling the ratio-normalized spectral strength via ρ with $\epsilon = 10^{-8}$ (Eq. 11). We use $\rho = 2 \times 10^{-2}$, and we additionally report an ablation over different ρ values in Sec. 5.3. For CFG, we



(a) Histogram of coupling flip-rate across mini-batches.

(b) Ablation on the SOT spectral strength ρ.

Fig. 3: Coupling disagreement and spectral-strength ablation. (on CSL-Daily) Left: Flip-rate is the fraction of rows whose dominant matched partner differs between pose-only OT coupling P_{pose} and SOT coupling P_{sot}; aggregated over N=300 minibatches uniformly subsampled during training at $\rho = 2 \times 10^{-2}$. **Right:** BLEU-4 as a function of ρ; performance peaks at $\rho = 2 \times 10^{-2}$. See the supplementary material for details on the flip-rate computation.

train with condition dropout by dropping the text condition with probability $p_{\text{uncond}} = 0.1$, and apply guidance at inference with scale $\omega = 4$.

5 Results

5.1 Quantitative Results

Table 1 compares our method with prior approaches on CSL-Daily and Phoenix-2014T. Pose accuracy is measured with the DTW-based metrics of SOKE [54], and BLEU-4 (B-4) is computed with a back-translation evaluator trained by us, so that all methods are scored under the same semantic evaluation. Our method obtains the best results on every metric across both datasets. The margin over the strongest baseline, SignFlow [18], is modest on the pose metrics (e.g., PA-JPE Body 4.49→4.42 on CSL-Daily and 4.47→4.32 on Phoenix-2014T) but substantial on the semantic one: BLEU-4 improves from 6.84 to 9.09 on CSL-Daily (+2.25) and from 8.91 to 9.78 on Phoenix-2014T (+0.87). This distribution of gains matches the design of our coupling strategy, which targets semantic alignment under guidance rather than raw joint accuracy: SOT preserves the pose fidelity of the CFM-OT baseline while clearly strengthening sentence-level semantic fidelity.

5.2 Qualitative Results

Figure 2 presents qualitative comparisons on the CSL-Daily and Phoenix-2014T test sets, respectively, showing ground-truth (GT) signer frames alongside avatar

Table 2: Back-translation results on CSL-Daily and Phoenix-2014T (standard split). *Guided* denotes classifier-free guided sampling at inference, while *Unguided* uses the same trained model without guidance ($\omega = 0$). *GT (upper bound)* reports back-translation scores when the evaluator is fed ground-truth motions. Best and second-best results within each dataset/sampling block (excluding GT) are in **bold** and underlined, respectively.

Method	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
CSL-Daily					*Unguided sampling*
GT (upper bound)	33.50	31.82	20.34	13.48	9.27
CFM-OT [18]	26.37	**26.39**	**15.41**	9.90	6.84
CFM-SOT	26.55	25.60	15.05	9.57	6.47
CFM-OT + CFG	25.63	25.20	14.45	9.11	6.21
CFM-SOT + CFG	**26.79**	26.09	15.30	**9.97**	**6.89**
CSL-Daily					*Guided sampling*
CFM-OT + CFG	30.46	28.89	18.09	12.10	8.44
CFM-SOT + CFG	**31.38**	**29.48**	**18.75**	**12.78**	**9.09**
Phoenix-2014T					*Unguided sampling*
GT (upper bound)	26.55	26.88	18.06	13.43	10.75
CFM-OT [18]	**23.43**	**23.99**	**15.48**	11.30	8.91
CFM-SOT	23.18	23.51	15.39	**11.40**	**9.07**
CFM-OT + CFG	22.15	23.22	15.01	11.06	8.80
CFM-SOT + CFG	22.37	22.57	14.15	10.28	8.13
Phoenix-2014T					*Guided sampling*
CFM-OT + CFG	23.26	23.44	15.50	11.40	8.97
CFM-SOT + CFG	**24.60**	**24.83**	**16.40**	**12.25**	**9.78**

renderings generated by SignFlow (CFM-OT) and our method. Across both datasets, our generations often show clearer signing dynamics: we observe clearer articulation changes over time and more distinct motion patterns in segments where the input sentence implies specific temporal structure (e.g., repetition and path extent). In addition, our results often produce hand configurations and spatial placements that appear closer to the GT realization, whereas SignFlow tends to yield smoother, more generic trajectories that can blur such dynamics-sensitive variations. These examples complement the quantitative gains and suggest that incorporating dynamics-aware coupling improves the consistency of conditional generation under guidance.

Table 3: Where to apply SOT under CFG-style masking (CSL-Daily, *unguided* sampling with $\omega = 0$). We apply SOT only to the masked or unmasked coupling step (masking rate 10%) and compare against unguided baselines.

Coupling configuration	**BLEU-4**
CFM-OT + CFG (unguided)	6.21
SOT on masked (10%) + OT on unmasked	6.68
SOT on unmasked + OT on masked (10%)	6.35
CFM-SOT + CFG (unguided)	**6.89**

5.3 Ablation Study

Effect of guidance. Table 2 reports back-translation performance, which we use throughout this section as a proxy for sentence-level semantic fidelity.

Guidance helps at inference. Applying CFG at inference improves semantic alignment on both datasets. On CSL-Daily, BLEU-4 rises from 6.84 for unguided CFM-OT to 8.44 for CFM-OT+CFG with guided sampling, and ROUGE-L from 26.37 to 30.46. Phoenix-2014T follows the same pattern with a smaller margin: guided sampling lifts CFM-OT+CFG from 8.80 to 8.97 BLEU-4 and from 22.15 to 23.26 ROUGE-L. Whatever the coupling, guidance makes the generated sequences adhere more closely to the input text.

Condition dropout hurts training under Euclidean coupling. Guidance at inference, however, requires condition dropout at training time, and this dropout has a cost. To measure it, we evaluate dropout-trained models with guidance switched off ($\omega = 0$). On CSL-Daily, CFM-OT+CFG evaluated this way falls below the plain CFM-OT baseline (BLEU-4 6.21 vs. 6.84; ROUGE-L 25.63 vs. 26.37): sharing parameters between masked and unmasked updates degrades the model when the coupling is purely pose-based. Replacing the Euclidean coupling with SOT recovers the loss, and CFM-SOT+CFG reaches 6.89 BLEU-4 and 26.79 ROUGE-L without guidance, ahead of both the degraded CFM-OT+CFG and the original baseline. This is what we would expect if SOT reduces dynamics-inconsistent pairings in masked updates, making the masked and unmasked training signals more compatible.

SOT and guidance compound. The two effects combine at inference. With guided sampling, CFM-SOT+CFG gives the best results on both datasets: on CSL-Daily it adds +0.65 BLEU-4 (8.44 $\rightarrow$ 9.09) and +0.92 ROUGE-L (30.46 $\rightarrow$ 31.38) over CFM-OT+CFG, and on Phoenix-2014T the margin is larger still, +0.81 BLEU-4 (8.97 $\rightarrow$ 9.78) and +1.34 ROUGE-L (23.26 $\rightarrow$ 24.60). Coupling by temporal spectral structure thus supplies supervision that guidance can exploit at inference, beyond what either mechanism achieves alone.

Dataset dependence under unguided sampling. Phoenix-2014T exhibits a different unguided pattern than CSL-Daily. SOT alone slightly improves BLEU-4 over CFM-OT (8.91 → 9.07), while condition-dropout training without guided inference can reduce performance (e.g., CFM-OT+CFG: 8.80; CFM-SOT+CFG: 8.13). We conjecture that the impact of masking and coupling depends on dataset-specific conditional variability, but the consistent improvements under *guided* inference across both datasets align with our main claim.

Sensitivity to spectral strength ρ. We ablate the spectral strength ρ, which controls the relative contribution of the DCT-based temporal descriptor in the SOT assignment cost via the ratio-normalized weight in Eq. (11). Intuitively, ρ acts as a knob that determines how strongly the spectral distance influences the OT assignment *relative* to the pose-space distance: larger ρ yields couplings that prioritize spectral similarity more heavily, while smaller ρ makes SOT closer to standard Euclidean OT. Figure 3b shows BLEU-4 on CSL-Daily as ρ varies. We observe a clear optimum at $\rho = 2 \times 10^{-2}$. When ρ is increased beyond this range (e.g., $\rho \geq 10^{-1}$), performance degrades, suggesting that over-weighting the spectral term can dominate the coupling and produce assignments that are less compatible with conditional generation. Conversely, when ρ is too small (e.g., 10^{-5}), the spectral term becomes weak, and the gain diminishes.

Evidence that SOT primarily benefits masked-branch coupling. To probe which branch benefits most from SOT under CFG-style condition dropout, we apply SOT only when computing the OT assignment for either the masked (unconditional) or unmasked (conditional) branch, while using Euclidean OT for the other branch. We evaluate with *unguided* sampling ($\omega = 0$) to isolate training effects from guided inference. Table 3 shows that applying SOT to the *masked* branch yields a larger improvement over the Euclidean baseline than applying it to the unmasked branch (BLEU-4 6.68 vs. 6.35, compared to 6.21 for CFM-OT+CFG). This suggests that coupling errors in masked updates contribute substantially to the degradation observed under shared-parameter dropout training, and that SOT mitigates this by producing more dynamics-consistent assignments. Applying SOT to both branches (CFM-SOT+CFG) achieves the best result (6.89), indicating that both branches benefit, with the masked branch being the dominant contributor in this setting.

Coupling evidence. To verify that the spectral term in SOT meaningfully affects minibatch OT matchings, we measure *coupling disagreement* between pose-only OT and SOT. Concretely, for the same minibatches we compute (i) an OT coupling using only the pose cost and (ii) an OT coupling using the full SOT cost, and quantify their difference using two complementary indicators: (1) the normalized ℓ_1 distance between the resulting transport plans, and (2) a *flip-rate* measuring how often the dominant matched partner changes. As expected, when $\rho = 0$ (pose OT), both metrics are 0. At our selected operating setting $\rho = 2{\times}10^{-2}$, we observe a consistent non-zero coupling shift, indicating that SOT

Table 4: Coupling disagreement induced by SOT (on CSL-Daily). Flip-rate measures how often the dominant matched partner changes between P_{pose} and P_{sot} (row-wise arg max transported mass). Δ_{plan} is the mean absolute difference between the two OT plans, normalized by B^2. Reported as mean±std over N=300 logged minibatches uniformly subsampled across training at $\rho = 2 \times 10^{-2}$.

ρ	Flip-rate (%) ↑	Δ_{plan} ($\times 10^{-5}$) ↑
0 (pose OT)	0.00 ± 0.00	0.00 ± 0.00
2×10^{-2}	$\mathbf{7.70 \pm 6.70}$	$\mathbf{3.77 \pm 3.27}$

alters minibatch pairings and therefore changes the OT-based supervision signal used in conditional flow matching. Figure 3a shows the distribution of flip-rate across N=300 minibatches uniformly subsampled during training: while some minibatches exhibit zero flips, the histogram contains substantial non-zero mass and a long tail, confirming that SOT can meaningfully modify OT matchings across training. Finally, Figure 3b ablates the spectral strength ρ and reveals a clear optimum at $\rho = 2 \times 10^{-2}$, consistent with the above evidence that moderate spectral weighting produces beneficial coupling changes. We observe consistent trends on Phoenix-2014T; corresponding ablations are reported in the supplementary material.

6 Conclusion

We studied flow-based text-to-sign language generation under a one-to-many setting and investigated how classifier-free guidance (CFG) and minibatch optimal transport (OT) coupling interact during training and sampling. While CFG substantially improves semantic alignment, we hypothesized that standard Euclidean OT coupling can produce dynamics-inconsistent assignments under condition dropout, reducing the compatibility between the conditional and unconditional signals that guidance combines at inference. To address this, we proposed Spectral Optimal Transport (SOT). Across experiments, SOT consistently amplifies the gains of CFG and yields the best quantitative and qualitative results. As future work, we plan to extend the pose-only setting by incorporating facial expressions, which are essential for fully expressive sign language but were outside the scope of this study, which focused on body and hand motion.

Acknowledgements

This work was supported in part by JST Moonshot R&D under Grant JPMJMS2011, JSPS KAKENHI Grant No. 26H02525, and ROIS NII Open Collaborative Research 2026-261S04-24183.

References

1. Akhter, I., Sheikh, Y., Khan, S., Kanade, T.: Nonrigid structure from motion in trajectory space. Advances in neural information processing systems **21** (2008)
2. Arkushin, R.S., Moryossef, A., Fried, O.: Ham2pose: Animating sign language notation into pose sequences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21046–21056 (2023)
3. Baltatzis, V., Potamias, R.A., Ververas, E., Sun, G., Deng, J., Zafeiriou, S.: Neural sign actors: A diffusion model for 3d sign language production from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1985–1995 (2024)
4. Berndt, D.J., Clifford, J.: Using dynamic time warping to find patterns in time series. In: Proceedings of the 3rd international conference on knowledge discovery and data mining. pp. 359–370 (1994)
5. Bosworth, R.G., Wright, C.E., Dobkins, K.R.: Analysis of the visual spatiotemporal properties of american sign language. Vision research **164**, 34–43 (2019)
6. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7784–7793 (2018)
7. Cheng, H.K., Schwing, A.: The curse of conditions: Analyzing and improving optimal transport for conditional flow-based generation. arXiv preprint arXiv:2503.10636 (2025)
8. Emmorey, K., Casey, S.: A comparison of spatial language in english & american sign language. Sign Language Studies **88**(1), 255–288 (1995)
9. Fan, W., Zheng, A.Y., Yeh, R.A., Liu, Z.: Cfg-zero*: Improved classifier-free guidance for flow matching models. arXiv preprint arXiv:2503.18886 (2025)
10. Fang, S., Sui, C., Zhang, X., Tian, Y.: Signdiff: Learning diffusion models for american sign language production. arXiv e-prints pp. arXiv–2308 (2023)
11. Feng, L., Wang, L., Hu, K., Kong, D., Wang, Z.: Text2sign diffusion: A generative approach for gloss-free sign language production. arXiv preprint arXiv:2509.10845 (2025)
12. Feng, R., Yu, C., Deng, W., Hu, P., Wu, T.: On the guidance of flow matching. arXiv preprint arXiv:2502.02150 (2025)
13. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
14. Huang, W., Pan, W., Zhao, Z., Tian, Q.: Towards fast and high-quality sign language production. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3172–3181 (2021)
15. Hwang, E.J., Cho, S., Lee, J., Park, J.C.: An efficient gloss-free sign language translation using spatial configurations and motion dynamics with llms. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 3901–3920 (2025)
16. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. Advances in Neural Information Processing Systems **36**, 20067–20079 (2023)
17. Khan, N., Wu, B., Ishi, C.T., Nakadai, K.: Multigau: Real time sign language generation using multimodal gated attention. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. pp. 149–160. Springer (2025)

18. Khan, N., Wu, B., Tan, S., Ishi, C.T., Nakadai, K.: Signflow: End-to-end sign language generation for one-to-many modeling using conditional flow matching. In: Proceedings of the 27th International Conference on Multimodal Interaction. pp. 173–180 (2025)
19. Kızıltepe, T., Taşyürek, S.M., Keles, H.Y.: Iterative latent refinement for robust non-autoregressive sign language production. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4883–4893 (2025)
20. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
21. Li, D., Rodriguez, C., Yu, X., Li, H.: 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. pp. 1459–1469 (2020)
22. Li, M., Chen, S., Zhang, Z., Xie, L., Tian, Q., Zhang, Y.: Skeleton-parted graph scattering networks for 3d human motion prediction. In: European conference on computer vision. pp. 18–36. Springer (2022)
23. Lin, C.Y., Och, F.J.: Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL-04). pp. 605–612 (2004)
24. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
25. Mao, W., Liu, M., Salzmann, M.: Generating smooth pose sequences for diverse human motion prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13309–13318 (2021)
26. Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9489–9497 (2019)
27. Müller, M., Jiang, Z., Moryossef, A., Gonzales, A.R., Ebling, S.: Considerations for meaningful sign language machine translation based on glosses. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 682–693 (2023)
28. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
29. Qi, F., Duan, Y., Zhang, H., Xu, C.: Signgen: End-to-end sign language video generation with latent diffusion. In: European Conference on Computer Vision. pp. 252–270. Springer (2024)
30. Ren, Z., Pan, Z., Zhou, X., Kang, L.: Diffusion motion: Generate text-guided 3d human motion by diffusion model. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
32. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
33. Sandler, W., Lillo-Martin, D.C.: Sign language and linguistic universals. Cambridge University Press (2006)

34. Saunders, B., Camgoz, N.C., Bowden, R.: Progressive transformers for end-to-end sign language production. In: European Conference on Computer Vision. pp. 687–705. Springer (2020)
35. Saunders, B., Camgoz, N.C., Bowden, R.: Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks. International journal of computer vision **129**(7), 2113–2135 (2021)
36. Saunders, B., Camgoz, N.C., Bowden, R.: 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. pp. 5141–5151 (2022)
37. Stokoe Jr, W.C.: Sign language structure: An outline of the visual communication systems of the american deaf. Journal of deaf studies and deaf education **10**(1), 3–37 (2005)
38. Stoll, S., Mustafa, A., Guillemaut, J.Y.: There and back again: 3d sign language generation from text using back-translation. In: 2022 International Conference on 3D Vision (3DV). pp. 187–196. IEEE (2022)
39. Tan, S., Miyazaki, T., Khan, N., Nakadai, K.: Improvement in sign language translation using text ctc alignment. In: Proceedings of the 31st International Conference on Computational Linguistics. pp. 3255–3266 (2025)
40. Tang, S., Hong, R., Guo, D., Wang, M.: 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 (2022)
41. Tang, S., Xue, F., Wu, J., Wang, S., Hong, R.: Gloss-driven conditional diffusion models for sign language production. ACM Transactions on Multimedia Computing, Communications and Applications **21**(4), 1–17 (2025)
42. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. The Eleventh International Conference on Learning Representations (2022)
43. Tian, S., Zheng, M., Liang, X.: Transfusion: A practical and effective transformer-based diffusion model for 3d human motion prediction. IEEE Robotics and Automation Letters **9**(7), 6232–6239 (2024)
44. Tong, A., Fatras, K., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Wolf, G., Bengio, Y.: Improving and generalizing flow-based generative models with minibatch optimal transport. arXiv preprint arXiv:2302.00482 (2023)
45. Walsh, H., Saunders, B., Bowden, R.: Select and reorder: A novel approach for neural sign language production. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 14531–14542 (2024)
46. Xie, P., Peng, T., Du, Y., Zhang, Q.: Sign language production with latent motion transformer. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3012–3022. IEEE Computer Society (2024)
47. Xie, P., Zhang, Q., Taiying, P., Tang, H., Du, Y., Li, Z.: G2p-ddm: Generating sign pose sequence from gloss sequence with discrete diffusion model. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6234–6242 (2024)
48. Xing, C., Mao, W., Liu, M.: Scene-aware human motion forecasting via mutual distance prediction. In: European Conference on Computer Vision. pp. 128–144. Springer (2024)
49. Yin, A., Li, H., Shen, K., Tang, S., Zhuang, Y.: T2s-gpt: Dynamic vector quantization for autoregressive sign language production from text. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3345–3356 (2024)

50. Yin, K., Moryossef, A., Hochgesang, J., Goldberg, Y., Alikhani, M.: Including signed languages in natural language processing. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 7347–7360 (2021)
51. Zhang, H., Shalev-Arkushin, R., Baltatzis, V., Gillis, C., Laput, G., Kushalnagar, R., Quandt, L.C., Findlater, L., Bedri, A., Lea, C.: Towards ai-driven sign language generation with non-manual markers. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. pp. 1–26 (2025)
52. Zheng, Q., Le, M., Shaul, N., Lipman, Y., Grover, A., Chen, R.T.: Guided flows for generative modeling and decision making. arXiv preprint arXiv:2311.13443 (2023)
53. Zhou, H., Zhou, W., Qi, W., Pu, J., Li, H.: Improving sign language translation with monolingual data by sign back-translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1316–1325 (2021)
54. Zuo, R., Potamias, R.A., Ververas, E., Deng, J., Zafeiriou, S.: Signs as tokens: A retrieval-enhanced multilingual sign language generator. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23806–23816 (2025)
55. Zuo, R., Wei, F., Chen, Z., Mak, B., Yang, J., Tong, X.: A simple baseline for spoken language to sign language translation with 3d avatars. In: European Conference on Computer Vision. pp. 36–54. Springer (2024)