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], Kazuhiro Nakadai[1]

[1] Institute of Science Tokyo [2] RIKEN Guardian Robot Project [3] ATR Hiroshi Ishiguro Laboratories [4] RIKEN Center for Biosystems Dynamics Research

nabeela@ra.sc.eng.isct.ac.jp

ECCV 2026 | MALMÖ | SEPT 8–12

MOCHA workshop

1. Introduction

- **Sign language generation (SLG)** maps spoken-language text to sign pose sequences; the goal is **semantic alignment**, the generated motion must convey the intended meaning.
- In sign language, **how a motion evolves over time** changes the meaning: the same handshape with a repeated vs. single movement, or a small vs. large path, is a **different** word as following [1]:

- **Motion dynamics are not style, they are part of the meaning.** Transferring the full semantics of a sentence therefore requires generating the right dynamics, not only the right poses. That is our aim.

2. Hypothesis 1: guidance strengthens semantics but does not respect dynamics

Unmasked update (text kept)
targets ~ p(x | y): concentrated dynamics

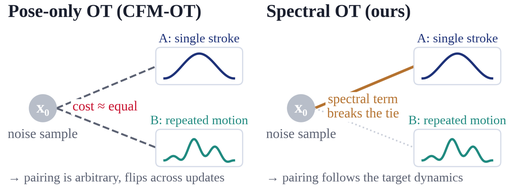
Masked update (text dropped)
targets ~ p(x): mixed dynamics

↓

One shared network vθ, one minibatch OT coupling
Euclidean pose cost is blind to motion dynamics

Consequence: CFG (Classifier free guidance) improves semantic alignment, but the pose-only coupling pairs noise with dynamics-incompatible targets, worst in the masked branch, so the masked and unmasked training signals conflict, and guidance combines two incompatible fields at inference.

3. Hypothesis 2: the coupling itself must account for motion dynamics



Spectral OT (SOT) re-weights the coupling cost toward temporal-frequency structure, so near-tied assignments are resolved consistently with the targets' dynamics, making the masked and unmasked signals compatible.

4. Contributions

1 **CFG helps flow-based SLG.** Empirical results shows that classifier-free guidance improves semantic alignment for text-to-sign flow matching.

2 **Spectral Optimal Transport (SOT).** Frequency-aware minibatch OT coupling adds a DCT-based spectral distance to the coupling cost so assignment respects motion dynamics.

3 **SOT × CFG compound.** SOT reduces the masked / unmasked mismatch and amplifies the gains from guidance: best on every metric on CSL-Daily [4] and Phoenix-2014T [5].

5. Spectral OT coupling

Noise batch $X_0 \sim \mathcal{N}(0, I)$ — **Target motions X_1 (+ text Y)**

Pose cost M_pose(i, j)
$|\mathrm{vec}(x_0^i) - \mathrm{vec}(x_1^j)|^2$ (as in CFM-OT)

Spectral cost M_spec(i, j) — NEW
temporal DCT $\psi(\cdot)$ with gain w_k, then $|\psi(x_0^i) - \psi(x_1^j)|^2$

Combined cost M = M_pose + λ_spec · M_spec
λ_spec = ρ · 𝔼[M_pose] / (𝔼[M_spec] + ε) — scale-free, single knob ρ

Hungarian assignment $\pi^* = \mathrm{argmin}_\pi \Sigma_i M(i, \pi(i))$

Coupled pairs $(x_0^i, x_1^{\pi(i)}, y^i)$ → straight CFM-OT paths
train vθ with condition dropout; guide at inference (ω = 4)

Drop-in: Same architecture, text encoder, schedule and optimizer as SignFlow; only the coupling cost changes. Since x_0 is isotropic, $\psi(x_0)$ is a dynamics-neutral probe , the dynamics enter from the **target** side.

6. Key equations

Temporal DCT descriptor with frequency gain (β)

$$\Psi_{k,d}(\mathbf{x}) = \sqrt{w_k} \sum_{t=0}^{T-1} x_{t,d} \cos\left(\tfrac{\pi}{T}\left(t+\tfrac{1}{2}\right)k\right), \qquad w_k = 1 + (\beta - 1)\,\tfrac{k}{T-1}$$

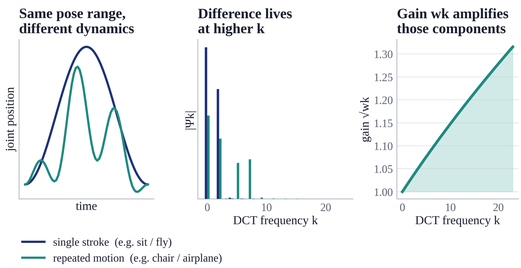
Coupling cost

$$M(i,j) = \underbrace{\left\|\mathrm{vec}(\mathbf{x}_0^i) - \mathrm{vec}(\mathbf{x}_1^j)\right\|_2^2}_{M_{\text{pose}}\ \text{(pose similarity)}} + \lambda_{\text{spec}} \underbrace{\left\|\psi(\mathbf{x}_0^i) - \psi(\mathbf{x}_1^j)\right\|_2^2}_{M_{\text{spec}}\ \text{(temporal dynamics)}}$$

Ratio-normalized weight and assignment

$$\lambda_{\text{spec}} = \rho \cdot \frac{\mathbb{E}[M_{\text{pose}}]}{\mathbb{E}[M_{\text{spec}}] + \epsilon}, \qquad \pi^\star = \arg\min_\pi \sum_{i=1}^{B} M\left(i, \pi(i)\right) \ \text{(Hungarian)}$$

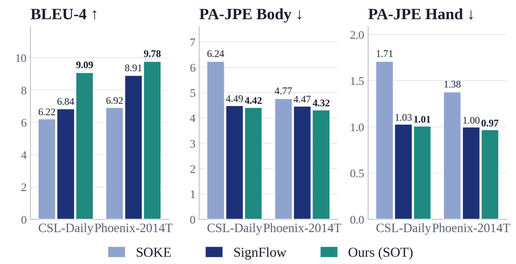
7. Why a spectral term?



Sign-motion energy sits at low frequencies, so a pose-only cost is decided by the slow, gross trajectory. **Repetition and articulation speed live at higher k**; the gain w_k raises their share, so targets that differ in dynamics become distinguishable during coupling.

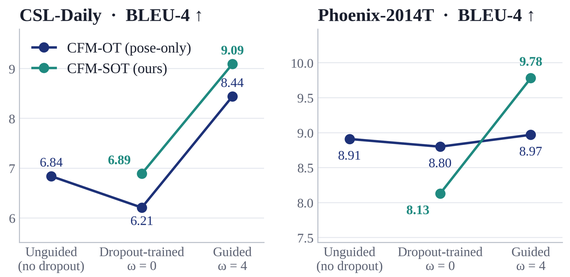
8. Quantitative Results

- **Semantics improve most.** BLEU-4 rises 6.84 → 9.09 on CSL-Daily (+2.25) and 8.91 → 9.78 on Phoenix-2014T (+0.87) over SignFlow [3].
- **Pose fidelity is preserved.** PA-JPE / JPE (body and hand) are equal or slightly better on both datasets, SOT does not trade accuracy for semantics.
- **Best on all 10 metrics** against six prior methods, with the same architecture, training schedule and evaluator.



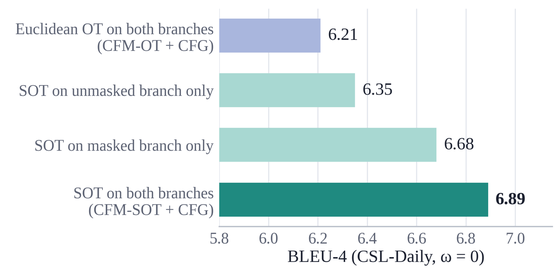
	CSL-Daily					Phoenix-2014T				
Method	**PA-JPE B↓**	**PA-JPE H↓**	**JPE B↓**	**JPE H↓**	**B-4↑**	**PA-JPE B↓**	**PA-JPE H↓**	**JPE B↓**	**JPE H↓**	**B-4↑**
SOKE[2]	6.24	1.71	7.38	9.68	6.22	4.77	1.38	6.04	7.72	6.92
SignFlow[3]	4.49	1.03	7.45	4.51	6.84	4.47	1.00	4.78	6.52	8.91
Ours (SOT)	**4.42**	**1.01**	**6.87**	**4.43**	**9.09**	**4.32**	**0.97**	**4.76**	**6.06**	**9.78**

9. Guidance and SOT combined



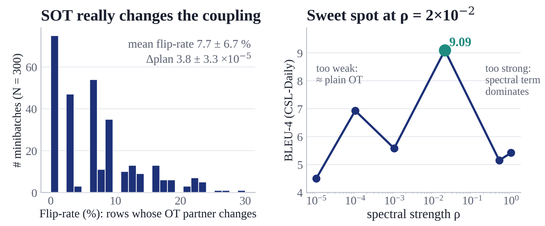
Condition dropout hurts pose-only OT (6.84 → 6.21 at ω = 0 on CSL-Daily); **SOT recovers the loss** (6.89) and, once guided, adds +0.65 / +0.81 BLEU-4 on top of CFM-OT + CFG.

10. Where does SOT help? The masked branch



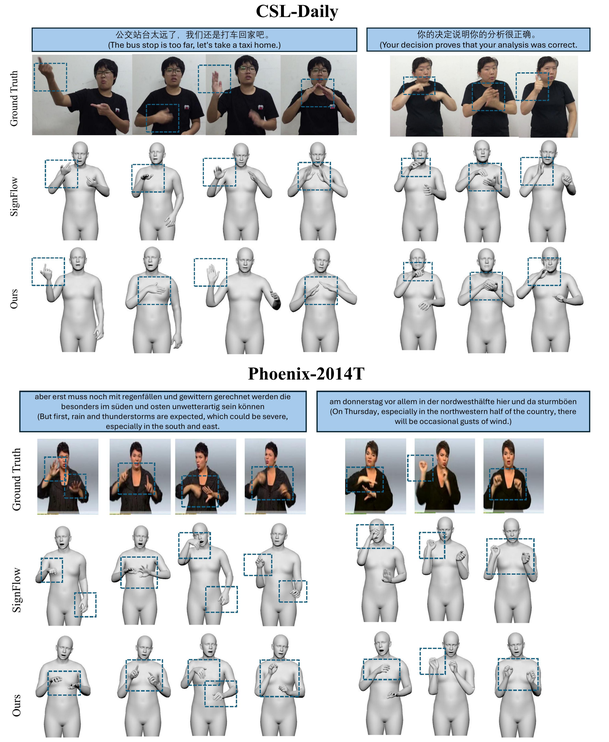
Applying SOT only to the masked (unconditional) coupling recovers most of the gain, coupling errors in masked updates are the dominant cause of the dropout degradation.

11. Does SOT change the coupling?



Flip-rate and plan difference vs. pose-only OT over N = 300 minibatches (CSL-Daily); both are exactly 0 at ρ = 0.

12. Qualitative: clearer signing dynamics



Dashed boxes mark hand configuration and placement; ours follows the GT articulation changes where SignFlow drifts to smoother, generic trajectories.

13. Take-away

- **Coupling by temporal spectral structure supplies supervision that guidance can exploit.** SOT consistently amplifies CFG on two benchmarks.
- **Cost:** one temporal DCT per minibatch and a single knob ρ; architecture, data and schedule untouched.
- **Caveat:** unguided behavior is dataset-dependent (Phoenix-2014T); the guided gains are consistent on both.
- **Next:** facial expressions and non-manual markers, beyond body and hand motion.

[1] Li, Dongxu, Cristian Rodriguez, Xin Yu, and Hongdong Li. "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.
[2] Zuo, Ronglai, Rolandos Alexandros Potamias, Evangelos Ververas, Jiankang Deng, and Stefanos Zafeiriou. "Signs as tokens: A retrieval-enhanced multilingual sign language generator." In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23806-23816. IEEE, 2025.
[3] Khan, Nabeela, Bowen Wu, Sihan Tan, Carlos Toshinori Ishi, and Kazuhiro Nakadai. "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.
[4]Zhou, Hao, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. "Improving sign language translation with monolingual data by sign back-translation." In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1316-1325. IEEE, 2021.[2] Zuo, Ronglai, Rolandos Alexandros Po
[5] Camgoz, Necati Cihan, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. "Neural sign language translation." In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7784-7793. Ieee, 2018.