NATURALNESS-GUIDED MANIFOLD FLOW MATCHING FOR SIGN LANGUAGE PRODUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sign Language Production (SLP) aims to generate sign motions from text. Conditional Flow Matching methods have achieved strong performance in SLP by constructing conditional paths that transform a source distribution into a target distribution. However, existing methods construct these paths via linear interpolation, whereas the rotational geometry of human joints confines valid joint rotations to a manifold embedded in Euclidean space. Consequently, linear interpolation between two sign motions leaves this manifold and ignores the motion distribution on it. In this paper, we revisit SLP from the perspective of manifold transport and propose a Naturalness-guided Manifold Flow Matching framework, termed **SignNMFlow**, which constructs conditional paths directly on the motion manifold by jointly considering geometric efficiency and the motion distribution. Specifically, we exploit the intrinsic geometry of the manifold and introduce a motion naturalness measure to characterize the motion distribution. By minimizing the kinetic energy under this measure, we learn a naturalness-guided interpolation that couples a closed-form geodesic, which provides geometrically efficient transport, with a learnable deviation that incorporates the motion distribution, thereby significantly improving the fidelity of generated sign motions. Extensive qualitative and quantitative evaluations demonstrate the effectiveness of this work.

1 INTRODUCTION

Sign language is a visual-spatial language that conveys rich linguistic information through precise hand shapes and coordinated body movements. Existing research has primarily focused on Sign Language Recognition (SLR) (Chen et al., 2022; Hu et al., 2023b; Zuo et al., 2023; Hu et al., 2023a) and Sign Language Translation (SLT) (Chen et al., 2022; Zhou et al., 2023; Gong et al., 2024; Wong et al., 2024; Jang et al., 2025), which aim to enable non-signers to understand sign language content. In contrast, Sign Language Production (SLP) (Baltatzis et al., 2024; Zuo et al., 2025; Tang et al., 2025; Low et al., 2026) has received relatively little attention. Unlike SLR and SLT, which decode existing sign language content, SLP must generate temporally coordinated motions that are semantically consistent with the text and visually natural.

Existing SLP methods span three families: autoregressive models (Saunders et al., 2020; Yin et al., 2024; Zuo et al., 2025), diffusion models (Baltatzis et al., 2024; Tang et al., 2025), and flow-based generative models (Khan et al., 2025; Low et al., 2026). Among these, Conditional Flow Matching (CFM) (Lipman et al., 2022) has demonstrated strong performance. CFM learns a velocity field that transports samples from a source distribution to the target distribution of sign motions, and anchors training on per-pair *conditional paths*. Existing flow matching-based SLP methods (Khan et al., 2025; Low et al., 2026) construct these paths by linear interpolation, thereby ignoring the rotational geometry of human joints and the distribution of sign motions. As shown in Figure 1(a), linear interpolation between two sign motions introduces invalid intermediate rotations. This invalidity has a geometric origin. A joint rotation is a geometric object, and rotation representations merely embed rotations into Euclidean space (Dünkel et al., 2024), so the valid rotations of a sign motion reside on a structured manifold (Sec. 2.1). As shown in Figure 1(b), a straight line between two sign motions in the ambient space therefore leaves the manifold, whereas its on-manifold counterpart, the geodesic, follows the intrinsic geometry. Moreover, linear interpolation depends only on the two endpoints and is independent of the distribution of sign motions. This geometric and distributional mismatch renders the learned velocity field inconsistent with the manifold.

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.

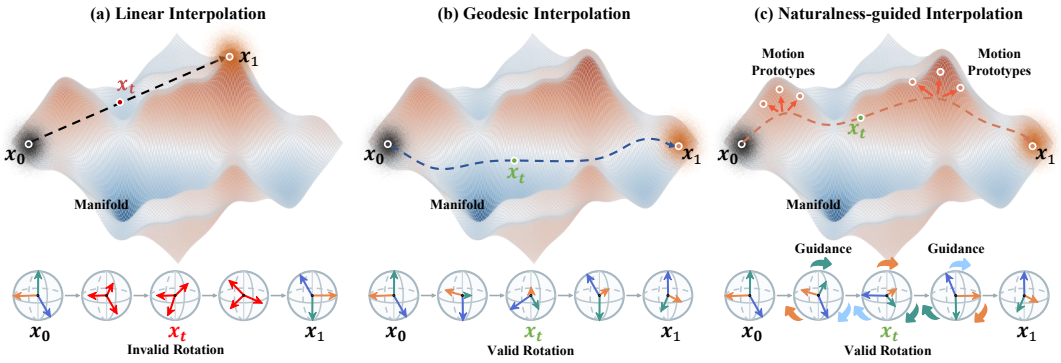

Figure 1: Here, x_0 and x_1 denote samples from the source and target distributions, and the color transition from blue to orange reflects the distribution of sign motions on the manifold. (a) Linear interpolation leaves the manifold, yielding invalid rotations. (b) Geodesic interpolation stays on the manifold but ignores the motion distribution. (c) Our naturalness-guided interpolation stays on the motion manifold, maintaining geometric efficiency while incorporating the motion distribution.

A better construction of conditional paths should be established for flow matching-based SLP. Although Sung et al. (2026) demonstrated that geodesics outperform linear interpolation in optimizing motion transitions between glosses[1], the improvement remains local to gloss transitions and leaves the geometric validity of the generative transport itself unexamined. We observe a geometric mismatch in existing flow matching-based SLP methods, where the conditional paths systematically deviate from the motion manifold while their construction ignores the motion distribution on it. Therefore, an ideal conditional path should stay on the manifold and ensure both geometric efficiency and consistency with the motion distribution, where geometric efficiency keeps the transport free of unnecessary detours and the motion distribution enhances motion fidelity.

To this end, we propose a Naturalness-guided Manifold Flow Matching framework, termed SignNMFlow. As shown in Figure 1(c), SignNMFlow constructs conditional paths directly on the motion manifold by jointly considering geometric efficiency and the motion distribution. Specifically, we first extract prototypical motions from sign motions and introduce a motion naturalness measure to characterize the distribution of sign motions on the manifold. Building on this, we construct the conditional paths through the naturalness-guided interpolation learned by minimizing the kinetic energy under this naturalness measure. This interpolation couples a geometric term, the closed-form geodesic of the manifold that provides geometrically efficient transport, with a naturalness-guided term, a learnable deviation from this geodesic that incorporates the motion distribution. In this way, SignNMFlow constructs geometrically efficient and motion-distribution-aware conditional paths, generating higher-fidelity sign motions. We summarize our **main contributions** as follows:

(1) We revisit SLP from the perspective of manifold transport and identify the geometric mismatch issue (Sec. 2.1) in existing flow matching-based SLP methods.
(2) We propose SignNMFlow (Sec. 3), a naturalness-guided manifold flow matching framework that constructs conditional paths directly on the motion manifold by jointly considering geometric efficiency and motion distribution.
(3) We introduce a motion naturalness measure (Sec. 3.2) that characterizes the distribution of sign motions on the manifold and learn the naturalness-guided interpolation by minimizing the kinetic energy under this measure (Sec. 3.3).
(4) Extensive qualitative and quantitative evaluations on the Phoenix-2014T, CSL-Daily, and How2Sign datasets demonstrate the effectiveness of SignNMFlow, which achieves hand DTW-PA-JPE of 1.07, 1.30, and 2.07, respectively.

2 Preliminaries

In this section, we identify the geometric mismatch between linear interpolation and the motion manifold of sign motions (Sec. 2.1), which motivates constructing conditional paths directly on the manifold. We then review flow matching (Sec. 2.2), which forms the basis of SignNMFlow.

[1]In linguistics, glosses are minimal lexical items that match the meaning of signs.

2.1 GEOMETRIC MISMATCH

Constrained by the rotational geometry of human joints, valid sign motions do not occupy the full Euclidean space but reside on a manifold embedded within it. Formally, the rotation of each joint in a sign motion is represented by a rotation matrix $Q \in SO(3)$, where $SO(3) = \{Q \in \mathbb{R}^{3\times3} \mid Q^\top Q = I, \det(Q) = 1\}$. Each frame of a sign motion therefore consists of a tuple of J rotations, and the space of all such tuples forms the J-fold product manifold $SO(3)^J$, which is embedded in the Euclidean space $\mathbb{R}^{9J}$. However, existing flow matching-based SLP methods (Khan et al., 2025; Low et al., 2026) construct conditional paths by linearly interpolating rotation representations in this space. We demonstrate the inherent incompatibility between linear interpolation and rotational geometry by directly analyzing $SO(3)$. Given a source rotation $R_0 \in SO(3)$ and a target rotation $R_1 \in SO(3)$, the conditional path constructed by linear interpolation between them is defined as:

$$R_t = (1-t)R_0 + tR_1, \quad t \in [0,1]. \tag{1}$$

Every $R \in SO(3)$ has orthonormal columns and hence satisfies $\|R\|_F = \sqrt{\mathrm{tr}(R^\top R)} = \sqrt{3}$. Expanding the squared Frobenius norm of the interpolated rotation via the Frobenius inner product $\langle\cdot,\cdot\rangle_F$ yields the closed form (see Appendix B.1):

$$\|R_t\|_F^2 = 3 - t(1-t)\,\|R_0 - R_1\|_F^2. \tag{2}$$

Whenever $R_0 \neq R_1$, we have $\|R_0 - R_1\|_F^2 > 0$, so for any $t \in (0,1)$, $\|R_t\|_F^2 < 3$. The interpolated state R_t therefore violates the orthogonality constraint that defines $SO(3)$, and linear interpolation inevitably produces intermediate states off the manifold. The transport adopted by existing flow matching-based SLP methods is thus misaligned with the intrinsic geometry of the motion manifold. We formalize this observation as Property 1.

Property 1. *Given two distinct rotations $R_0, R_1 \in SO(3)$, the linear interpolation $R_t = (1-t)R_0 + tR_1$, $t \in [0,1]$, produces intermediate states $R_t \notin SO(3)$ for all $t \in (0,1)$.*

In practice, joint rotations are commonly parameterized using the 6D rotation representation (Zhou et al., 2019), unit quaternions, or Euler angles. Across these representations, linear interpolation fails to respect the underlying rotation geometry. This mismatch motivates constructing conditional paths directly on the motion manifold. For further details, please refer to Appendix B.

2.2 FLOW MATCHING

Flow matching (FM) (Albergo & Vanden-Eijnden, 2022) learns a velocity field that transports samples from a source distribution p_0 to a target distribution p_1 along a predefined probability path p_t, $t \in [0,1]$. Unlike score-based diffusion (Sohl-Dickstein et al., 2015; Ho et al., 2020), which learns the score function of a stochastic process, FM directly minimizes the discrepancy between a learnable vector field $v_t^\theta(x)$, given by a neural network with parameters θ, and a target velocity field $u_t(x)$:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{t\sim\mathcal{U}[0,1],\, x\sim p_t}\left[\|v_t^\theta(x) - u_t(x)\|^2\right]. \tag{3}$$

However, the target field $u_t(x)$ requires marginalizing over the full data distribution and is thus intractable in practice. Conditional Flow Matching (CFM) (Lipman et al., 2022) circumvents this issue by conditioning on individual source-target pairs. Given a source sample $x_0 \sim p_0$ and a target sample $x_1 \sim p_1$, one defines a conditional probability path $p_t(x|x_0,x_1)$ and a conditional velocity field $u_t(x|x_0,x_1)$. Common constructions include linear interpolation in Euclidean space (Lipman et al., 2022) and geodesic interpolation on manifolds (Chen & Lipman, 2024). The CFM training objective can be stated as:

$$\mathcal{L}_{\mathrm{CFM}} = \mathbb{E}_{t\sim\mathcal{U}[0,1],\, x_0\sim p_0,\, x_1\sim p_1}\left[\|v_t^\theta(x_t) - u_t(x_t|x_0,x_1)\|^2\right], \tag{4}$$

where x_t is a sample from the conditional path at time t. This conditional objective is tractable and yields the same parameter gradients as $\mathcal{L}_{\mathrm{FM}}$, enabling efficient training through per-sample, path-based supervision.

3 NATURALNESS-GUIDED MANIFOLD FLOW MATCHING

We introduce NATURALNESS-GUIDED MANIFOLD FLOW MATCHING (SignNMFlow), which constructs conditional paths directly on the motion manifold by jointly considering geometric efficiency

and the motion distribution. SignNMFlow introduces a motion naturalness measure that characterizes the distribution of sign motions on the manifold and learns a naturalness-guided interpolation by minimizing the kinetic energy under this measure.

We organize this section as follows. In Sec. 3.1, we present a manifold-based representation of sign motions, capturing the intrinsic rotational geometry of human joints. In Sec. 3.2, we define the motion naturalness measure. In Sec. 3.3, we construct the naturalness-guided interpolation by minimizing the kinetic energy under this measure and train the sign motion generation network along the learned paths.

3.1 Motion Representation

Existing methods (Khan et al., 2025; Low et al., 2026) typically represent sign motions with the 6D rotation representation (Zhou et al., 2019) and treat its coordinates as unconstrained Euclidean variables. Consequently, linear interpolation drives intermediate states off the manifold (Sec. 2.1). To construct conditional paths directly on the motion manifold, we represent a sign motion x as J unit quaternions:

$$x = (q_1, \dots, q_J) \in \mathcal{M} = (\mathbb{S}^3)^J, \tag{5}$$

where J denotes the number of joints, and $\mathcal{M}$ is the J-fold product manifold of unit hyperspheres $\mathbb{S}^3$. Since q and $-q$ represent the same rotation, we remove this sign ambiguity in Appendix C.1. A sign motion sequence with T frames is represented as $\mathcal{X} = (x^1, \dots, x^T)$. For notational clarity, we omit the frame index and develop the geometry on a single sign motion x in the remainder of this paper. All definitions below extend to $\mathcal{X}$ by applying the same maps to each frame. For any point $x \in \mathcal{M}$, let $T_x\mathcal{M}$ denote the tangent space at x. The exponential map and the logarithmic map are defined as:

$$\mathrm{Exp}_x : T_x\mathcal{M} \to \mathcal{M}, \qquad \mathrm{Log}_x : \mathcal{M} \to T_x\mathcal{M}, \tag{6}$$

where Exp_x maps tangent vectors onto the manifold and Log_x is its local inverse.

3.2 Motion Naturalness Measure

A desirable conditional path on $\mathcal{M}$ should jointly consider geometric efficiency and the motion distribution. The former follows from the intrinsic geometry of $\mathcal{M}$, which Sec. 3.3 exploits to construct the naturalness-guided interpolation. The latter requires an explicit characterization of the motion distribution on $\mathcal{M}$. To this end, we introduce a motion naturalness measure.

We first extract prototypical motions from the training sign motions via K-means clustering (MacQueen, 1967), obtaining K prototypes $\{c_k\}_{k=1}^K$ with frequency weights $\{w_k\}_{k=1}^K$, where w_k reflects the occurrence frequency of the rotations assigned to prototype c_k. The detailed procedure of the K-means-based prototype extraction is provided in Appendix C.2. Given a motion $x \in \mathcal{M}$, we characterize the distribution of the prototypes around x through a kernel-weighted coefficient $\tilde{w}_k(x)$ and a local scatter matrix $S(x)$:

$$\tilde{w}_k(x) = w_k \cdot \exp\left(-\frac{\|\delta_k\|^2}{2\sigma^2}\right), \quad S(x) = \sum_{k=1}^{K} \tilde{w}_k(x) \cdot \delta_k \delta_k^\top, \tag{7}$$

where σ is the kernel bandwidth and $\delta_k = \mathrm{Log}_x(c_k) \in T_x\mathcal{M}$ is the tangent vector from x to prototype c_k. The kernel-weighted coefficient $\tilde{w}_k(x)$ quantifies the contribution of prototype c_k to the motion distribution around x, combining its global frequency weight w_k with a Gaussian proximity factor. The scatter matrix $S(x)$ complements this distributional information with the local dispersion of the prototypes around x, aggregating the outer products $\delta_k\delta_k^\top$ weighted by $\tilde{w}_k(x)$. We define the naturalness measure as:

$$G(x) = m(x)\big(S(x) + \rho\mathbf{I}\big)^{-1}, \tag{8}$$

where $m(x) = \sum_{k=1}^K \tilde{w}_k(x)$ denotes the aggregate kernel mass at x, and $\rho = 1e-6$ is a regularization constant. We establish the validity of this measure in Appendix C.3. The point-wise measure $G(x)$ alone cannot evaluate an entire conditional path. We therefore lift it from points to paths, incorporating the motion distribution into path construction. For a path $\gamma : [0,1] \to \mathcal{M}$ with $\gamma_0 = x_0$ and $\gamma_1 = x_1$, we define its length $\mathcal{N}(\gamma)$ under the naturalness measure as:

$$\mathcal{N}(\gamma) = \int_0^1 \|\dot{\gamma}_t\|_{G(\gamma_t)}\, dt, \quad \|\dot{\gamma}_t\|_{G(\gamma_t)} := \sqrt{\langle \dot{\gamma}_t, G(\gamma_t)^{-1}\dot{\gamma}_t \rangle}, \tag{9}$$

Algorithm 1 Training procedure of motion naturalness learning

Require: motion manifold $\mathcal{M}$, source and target distributions p_0, p_1, initialized correction network ϵ_θ, naturalness measure $G(\cdot)$

while Training **do**
  Sample $x_0 \sim p_0$, $x_1 \sim p_1$, and $t \sim \mathcal{U}[0,1]$
  $\xi = \mathrm{Log}_{x_0}(x_1)$, $\zeta(t) = \epsilon_\theta(t, x_0, x_1)$, $\omega(t) = t\,\xi + t(1-t)\,\zeta(t)$
  $x_{t,\theta} = \mathrm{Exp}_{x_0}(t\mathrm{Log}_{x_0}(x_1) + t(1-t)\epsilon_\theta(t, x_0, x_1))$ ▷ Eq. equation 11
  $\dot{x}_{t,\theta} = \left(D\mathrm{Exp}_{x_0}\right)_{\omega(t)} \left[\xi + (1-2t)\,\zeta(t) + t(1-t)\,\frac{\partial}{\partial t}\epsilon_\theta(t, x_0, x_1)\right]$
  $\ell(\theta) \leftarrow (\dot{x}_{t,\theta})^\top G(x_{t,\theta})^{-1}\dot{x}_{t,\theta}$ ▷ Estimate of $\mathcal{L}_G(\theta)$ in Eq. equation 12
  Update θ using gradient $\nabla_\theta \ell(\theta)$
end while
return naturalness-guided transport paths parameterized by ϵ_θ

where $\dot{\gamma}_t$ is the velocity of the path and $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product.

3.3 Motion Naturalness Learning

With the motion representation (Sec. 3.1) and the motion naturalness measure (Sec. 3.2) established on $\mathcal{M}$, we construct conditional paths that secure geometric efficiency through the intrinsic geometry of $\mathcal{M}$, and incorporate the motion distribution by minimizing the kinetic energy under this measure. We then train the sign motion generation network along the learned paths.

Learning the naturalness-guided interpolation. Given a source sample $x_0 \sim p_0$ and a target sign motion sample $x_1 \sim p_1$, an ideal conditional path $\gamma : [0,1] \to \mathcal{M}$ should satisfy the boundary conditions $\gamma_0 = x_0$ and $\gamma_1 = x_1$. In differential geometry, the geodesic follows the intrinsic geometry of $\mathcal{M}$ and provides the shortest connection between endpoints. However, the geodesic depends solely on the geometry of $\mathcal{M}$ and ignores the motion distribution. To incorporate the motion distribution, we define the naturalness-guided interpolation as the conditional path that minimizes the kinetic energy under the naturalness measure:

$$\gamma^\star = \underset{\gamma:\, \gamma_0 = x_0,\, \gamma_1 = x_1}{\arg\min} \varphi(\gamma), \quad \varphi(\gamma) := \mathbb{E}_{t\sim\mathcal{U}[0,1]}\left[\dot{\gamma}_t^\top G(\gamma_t)^{-1}\dot{\gamma}_t\right], \tag{10}$$

where φ is the kinetic energy under the naturalness measure, i.e., the energy form of $\mathcal{N}(\gamma)$ in Eq. 9. If the naturalness measure were uniform, the minimizer of φ would be exactly the geodesic, so any deviation of $\gamma_t^\star$ from the geodesic is driven by the non-uniform motion distribution. Since the exact minimizer over all admissible paths is intractable, we parameterize the interpolation around the geodesic with a learnable correction network ϵ_θ, whose output is orthogonally projected onto the tangent space $T_{x_0}\mathcal{M}$ (the projection operator and the exactness guarantee are provided in Appendix C.4.2):

$$x_{t,\theta} = \mathrm{Exp}_{x_0}(\underbrace{t \cdot \mathrm{Log}_{x_0}(x_1)}_{\text{Geometric}} + \underbrace{t(1-t) \cdot \epsilon_\theta(t, x_0, x_1)}_{\text{Naturalness}}), \quad t \in [0,1], \tag{11}$$

where the geometric term $t \cdot \mathrm{Log}_{x_0}(x_1)$ is the closed-form geodesic that provides geometrically efficient transport, and the naturalness term $t(1-t)\cdot\epsilon_\theta(\cdot)$ is a learnable deviation from it, trained by minimizing φ. The factor $t(1-t)$ vanishes at both endpoints, so the boundary conditions hold for any ϵ_θ. Minimizing $\varphi(\gamma)$ over the parameterized path $x_{t,\theta}$ yields our training objective (the training procedure is summarized in Algorithm 1):

$$\mathcal{L}_G(\theta) := \mathbb{E}_{t\sim\mathcal{U}[0,1],\, x_0\sim p_0,\, x_1\sim p_1}[\dot{x}_{t,\theta}^\top G(x_{t,\theta})^{-1}\dot{x}_{t,\theta}], \tag{12}$$

where $\dot{x}_{t,\theta}$ denotes the velocity of $x_{t,\theta}$ at time t. In practice, we compute $\dot{x}_{t,\theta}$ by automatic differentiation. The chain-rule expression in Algorithm 1 is its analytic equivalent.

Training generation along the learned paths. With the naturalness-guided interpolation $x_{t,\theta}$, we freeze the correction network ϵ_θ and train the sign motion generation network ϕ along this path. Rather than directly regressing the conditional velocity field as in standard CFM Eq. 4, we adopt a

Table 1: Comparisons with state-of-the-art sign language production (text-to-sign) methods. Best mean values are in **bold** and second-best are underlined.

Methods	Phoenix-2014T					CSL-Daily				
	DTW-JPE		DTW-PA-JPE		B-T	DTW-JPE		DTW-PA-JPE		B-T
	Body	Hand	Body	Hand	BLEU-4	Body	Hand	Body	Hand	BLEU-4
Pr. Tr. (Saunders et al., 2020)	15.01	31.77	13.67	11.95	4.94	16.30	32.63	15.98	12.91	3.07
T2Mesh (Stoll et al., 2022)	14.04	31.64	13.48	12.06	5.81	13.76	30.37	13.47	12.10	5.11
T2S-GPT (Yin et al., 2024)	11.65	19.09	10.38	6.47	9.06	12.32	15.43	11.94	5.93	8.94
S-MGPT (Jiang et al., 2023)	10.42	9.08	9.45	3.41	9.68	11.58	11.31	10.81	3.78	8.82
SignFlow (Khan et al., 2025)	5.40	6.21	4.56	1.20	–	7.56	8.82	6.43	1.65	–
SOKE (Zuo et al., 2025)	6.04	7.72	4.77	1.38	11.87	7.38	9.68	6.24	1.71	11.30
SignSparK (Low et al., 2026)	4.40	7.10	5.24	1.52	–	6.27	10.63	7.26	2.00	–
SignNMFlow	**4.37**	**5.57**	**3.85**	**1.07**	**16.15**	**5.73**	**7.43**	**5.10**	**1.30**	15.14
SignNMFlow-M	4.42	5.75	3.91	1.08	14.45	5.94	7.73	5.31	1.33	**15.23**

target-prediction parameterization. The generation network ϕ takes the intermediate state x_t along the learned path, the paired expression state e_t, the time step t, and the text condition c as input, and predicts the raw target motion $\tilde{x}_1$ together with the target expression $\hat{e}_1$. The motion prediction is normalized per joint onto $\mathcal{M}$ through the radial projection $\hat{x}_1^j = \tilde{x}_1^j / \|\tilde{x}_1^j\|$, $j = 1, \ldots, J$. Formally, the training objective for the motion channel is

$$\mathcal{L}_{\mathrm{NMF}}(\phi) = \mathbb{E}_{t\sim\mathcal{U}[0,1],\, x_0\sim p_0,\, x_1\sim p_1}\left[\|x_1 - \hat{x}_1\|^2\right]. \tag{13}$$

The facial expression channel is trained in Euclidean space under the same objective. At sampling time, each solver step re-predicts the target $\hat{x}_1$ from the current state, re-evaluates the correction with this prediction, and advances the state to the next path point of Eq. 11. The complete procedures are given in Appendix C.5. The transport underlying generation is thus both geometrically efficient and aware of the motion distribution, yielding higher-fidelity sign motions.

Geometric efficiency and naturalness on equal footing. Geometric efficiency and naturalness differ in origin. The former arises from the intrinsic geometry of $\mathcal{M}$ and enters the path through the fixed geodesic, whereas the latter arises from the motion distribution and enters the path through the learnable deviation optimized by the kinetic energy objective. Naturalness itself is defined in the same geometric terms, with both the scatter matrix $S(x)$ and the kinetic energy expressed in tangent spaces via the logarithmic map. In the learned path, the geodesic remains the fixed reference while the naturalness term absorbs the deviation induced by the motion distribution, so neither consideration is obtained at the expense of the other.

4 Experiments

4.1 Experimental Settings

Datasets. Following existing methods (Zuo et al., 2025; Low et al., 2026), we evaluate our method on three sign language datasets: Phoenix-2014T (Camgoz et al., 2018), CSL-Daily (Zhou et al., 2021), and How2Sign (Duarte et al., 2021), covering German, Chinese, and American sign language, respectively. For sign motion extraction, we adopt SMPL-X (Pavlakos et al., 2019) parameters, which are sourced from NSA (Baltatzis et al., 2024) for How2Sign and from SOKE (Zuo et al., 2025) for Phoenix-2014T and CSL-Daily. SMPL-X represents each sign motion with joint rotations and facial expression parameters. Our manifold construction applies to the joint rotations, while the expression parameters are generated alongside them in Euclidean space under the same target-prediction objective as Eq. 13 (Appendix C.5). SignNMFlow is trained on a single dataset, and its multilingual version, SignNMFlow-M, is trained on the combination of all three, following the training setup of SOKE (Zuo et al., 2025).

Evaluation Metrics. To evaluate our performance, we follow SOKE (Zuo et al., 2025), using Dynamic Time Warping on the joint positions (DTW-JPE) and its Procrustes-aligned version (DTW-PA-JPE) to assess the accuracy of the sign motions. Furthermore, we used BLEU-4 scores calculated through Back-Translation to assess the semantic intelligibility of the generated sign motions. To ensure fairness, our Back-Translation model is the same as SOKE.

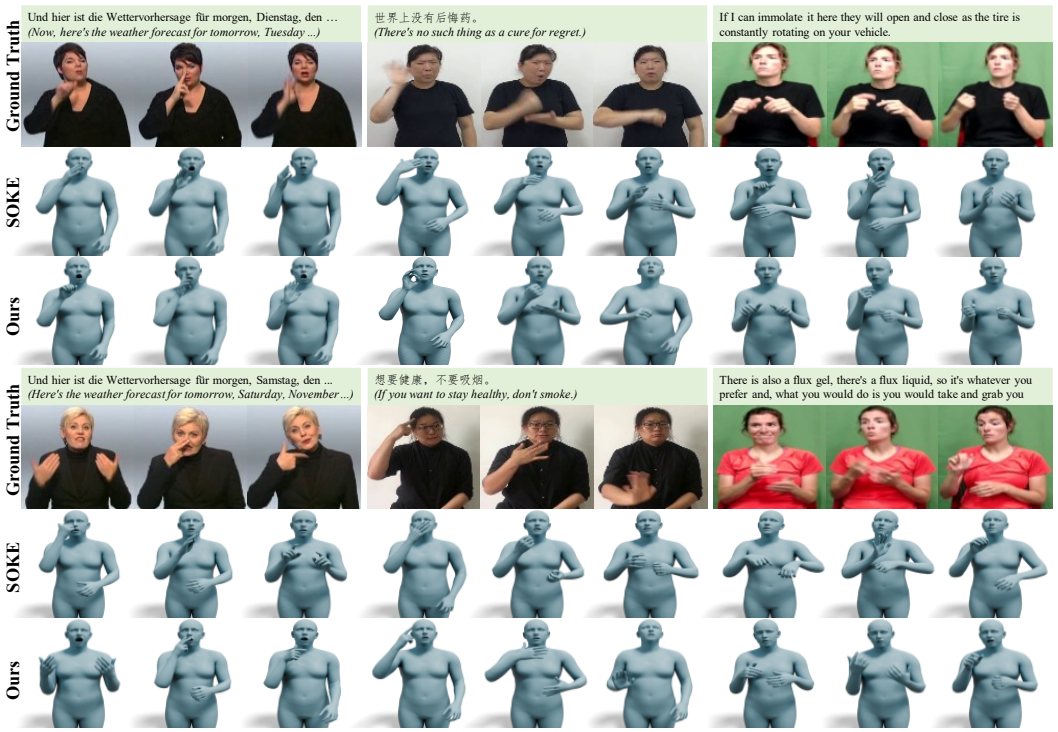

Figure 2: Qualitative comparisons of generated sign motions between our method with the SOKE, on the test sets of Phoenix-2014T(left), CSL-Daily(middle), and How2Sign(right).

Implementation Details. For the correction network ϵ_θ, we empirically use a 3 layers MLP with a hidden dimension of 512. We train the correction network ϵ_θ with a batch size of 8 for 100 epochs, using the AdamW optimizer (Loshchilov & Hutter, 2017) and a cosine learning rate scheduler starting at 2e-4. For the text encoder, we use the pre-trained mBART-large-cc25 (Liu et al., 2020). For the sign motion generation network ϕ, we use a transformer decoder network with the 8 decoder layers containing a separate cross-attention head for text. The timestep t is passed to the decoder layers through Stylization block (Zhang et al., 2024) after every self-attention, cross-attention, feed-forward layer. We train the denoising network with a batch size of 64 per GPU for 300 epochs, employing the same optimizer settings as the neural network ϵ_θ. All models are trained on 4 NVIDIA A40 GPUs. Further details are provided in Appendix D.

4.2 Comparison with State-of-the-Art Methods

Quantitative Comparison. In Table 1 and Table 2, we compare SignNMFlow with state-of-the-art gloss-free SLP methods, focusing on three representative methods: SignFlow (Khan et al., 2025), SOKE (Zuo et al., 2025), and SignSparK (Low et al., 2026). SOKE leverages a pre-trained language model with retrieval enhancements to improve SLP semantics; however, SignNMFlow achieves superior semantic fidelity despite these augmentations. SignFlow and SignSparK are both conditional flow matching frameworks that construct conditional paths by linear interpolation in Euclidean space, with SignFlow using optimal transport and SignSparK adopting sparse keyframes-based learning. Notably, SignNMFlow consistently outperforms both, achieving significantly lower hand DTW-PA-JPE errors of 1.07, 1.30, and 2.07 on Phoenix-2014T, CSL-Daily,

Table 2: Comparisons with state-of-the-art sign language production methods (text-to-sign) on How2Sign Dataset.

Methods	DTW-JPE		DTW-PA-JPE		B-T
	Body	Hand	Body	Hand	BLEU-4
Pr.Tr.	14.74	30.17	14.15	11.57	2.75
T2Mesh	15.50	32.97	13.99	13.47	7.51
T2S-GPT	12.65	18.44	11.48	6.39	11.20
S-MGPT	12.41	13.74	11.23	4.39	11.45
SignFlow	7.98	10.52	6.92	2.27	–
SOKE	7.75	10.08	6.82	2.35	14.48
SignSparK	**6.30**	11.43	7.26	2.72	–
SignNMFlow	6.76	**9.72**	6.01	2.07	16.88
SignNMFlow-M	6.74	9.77	**5.99**	**2.06**	**17.49**



Figure 3: Qualitative ablation for GE and NG.

Table 3: Ablation results of Motion Representation (MR), Geometric Efficiency (GE) and Naturalness-Guided (NG).

Modules			DTW-JPE↓			DTW-PA-JPE↓		
MR	GE	NG	All	Body	Hand	All	Body	Hand
×	×	×	20.08	6.70	9.25	10.78	6.07	1.68
✓	×	×	21.25	6.84	9.73	11.26	6.24	1.80
✓	✓	×	18.31	6.14	8.05	9.58	5.51	1.41
✓	✓	✓	**16.86**	**5.73**	**7.43**	**8.92**	**5.10**	**1.30**

Table 4: Ablation results of Measure.

Parameters		DTW-JPE↓			DTW-PA-JPE↓		
		All	Body	Hand	All	Body	Hand
K	250	17.17	5.83	7.49	**8.75**	5.19	1.32
	500	**16.86**	**5.73**	**7.43**	8.92	**5.10**	1.30
	750	17.32	5.87	7.53	8.80	5.20	**1.29**
σ	0.25	19.41	6.47	8.71	10.03	5.84	1.57
	0.50	**16.86**	**5.73**	**7.43**	8.92	**5.10**	**1.30**
	0.75	17.58	5.95	7.57	**8.80**	5.28	1.32

and How2Sign, respectively. Meanwhile, the multilingual version of SignNMFlow, SignNMFlow-M, has also demonstrated strong competitiveness. These results demonstrate the advantage of constructing conditional paths directly on the motion manifold.

Qualitative Comparison. In Figure 2, we present qualitative comparisons between SignNMFlow and SOKE (Zuo et al., 2025). Our method can continuously generate more realistic and expressive sign motions and significantly improve the fidelity of gestures. The advantage of SignNMFlow lies in its ability to construct conditional paths directly on the motion manifold, thereby more effectively preserving the intrinsic geometric structure of sign motions. Unlike SOKE, which may produce intermediate motion deviations, SignNMFlow constructs conditional paths through the naturalness-guided interpolation, generating sign motions that are significantly closer to the ground truth and demonstrating higher pose fidelity and spatial consistency. This highlights the effectiveness of SignNMFlow in capturing complex sign motion dynamics and generating high-fidelity sign motions.

4.3 Ablation Study

In this subsection, we present ablation results to verify the effectiveness of SignNMFlow. All results are evaluated on CSL-Daily, while Phoenix-2014T and How2Sign are not used for ablations.

Geometric Efficiency and Naturalness-Guided Analysis. In Table 3, we present the ablation results for Geometric Efficiency (GE) and Naturalness-Guided (NG). All variants are flow-matching SLP models sharing the same *target-prediction* parameterization. The baseline (first row) uses the common 6D rotation representation and constructs its conditional paths by linear interpolation in Euclidean space. Switching to the quaternion-based Motion Representation (MR) alone slightly raises DTW-JPE All error from 20.08 to 21.25. Linear interpolation treats quaternion components as flat coordinates rather than points on the rotation manifold, so the representation change alone brings no gain under this path geometry. GE removes this mismatch by constructing conditional paths as geodesics on the manifold instead of by linear interpolation. Paired with MR, GE lowers DTW-JPE All error from 21.25 to 18.31 and DTW-PA-JPE All error from 11.26 to 9.58, both below the 6D baseline (20.08 and 10.78). The quaternion representation becomes superior only once the path geometry agrees with the manifold on which rotations lie. In addition, NG considers the motion distribution on the manifold. It brings further gains, reducing the two metrics to 16.86 and 8.92, the best results across all variants and 16.04% and 17.25% below the baseline, with consistent improvements on body and hand joints. Furthermore, we perform a qualitative comparison in Figure 3 by removing NG (w/o NG), and further removing both NG and GE (w/o GE NG). The figure demonstrates that NG and GE effectively improve the quality and fidelity of generated motions.

Measure Configuration. In Table 4, we present the ablation results regarding the parameter configurations for the motion naturalness measure, where K is the number of motion prototypes and σ is the kernel bandwidth in Eq. 7. For K, performance peaks around 500 and mildly degrades at 750. Too few prototypes underfit the motion distribution, while too many over-fragment it into small

clusters whose frequency weights become unreliable. For σ, the effect is more pronounced. With $\sigma = 0.25$, the kernel weights $\tilde{w}_k(x)$ concentrate on the single nearest prototype, so $G(x)$ tracks local prototype fluctuations rather than the overall distribution, and the DTW-JPE All error rises from 16.86 to 19.41. With $\sigma = 0.75$, the kernel over-smooths the prototypes, so the weights become nearly uniform and $G(x)$ no longer reflects the non-uniform motion distribution. The interpolation accordingly loses part of its distributional guidance, and the DTW-JPE All error rises from 16.86 to 17.58, a milder degradation than at $\sigma = 0.25$. We therefore use $K = 500$ and $\sigma = 0.50$.

Training Objective. In Table 5, we compare *velocity-prediction* and *target-prediction* parameterizations of the training objective under both the CFM baseline and our method. For the CFM model, target prediction reduces DTW-JPE All error from 21.62 to 20.08 and DTW-PA-JPE All error from 11.33 to 10.78. By contrast, our method reduces these two metrics from 18.76 to 16.86 and from 9.52 to 8.92, respectively, with consistent performance improvements on both body and hand joints. The advantage stems from the regression target itself. Velocity prediction regresses the difference between the sampled noise and the data, and the scale of this target varies across samples. Target prediction instead regresses the ground-truth motion directly, and its supervision signal remains well scaled across timesteps.

Table 5: Ablation results of training objectives.

Method	Prediction	DTW-JPE↓			DTW-PA-JPE↓		
		All	Body	Hand	All	Body	Hand
CFM	Velocity	21.62	7.43	9.49	11.33	6.64	1.79
	Target	**20.08**	**6.70**	**9.25**	**10.78**	**6.07**	**1.68**
Ours	Velocity	18.76	6.44	8.13	9.52	5.69	1.44
	Target	**16.86**	**5.73**	**7.43**	**8.92**	**5.10**	**1.30**

4.4 User Study

In addition to objective metrics, we invited 5 professional CSL signers to participate in a user study. These signers made a forced choice based on the semantic alignment between the generated sign motions and text annotations, together with the fidelity of the motion. Higher user preference scores are associated with better semantic alignment and higher-fidelity motion. Specifically, we provided 18 generated sign motions from the SOTA baselines (SignFlow and SOKE) and SignNMFlow. The order of the sign motions was randomly shuffled to prevent potential bias. As shown in Figure 4, SignNMFlow was preferred in 81.11% of the trials, significantly outperforming SOKE (7.78%) and SignFlow (11.11%). These results are consistent with qualitative and quantitative evaluations, demonstrating the improvements of SignNMFlow in semantic alignment and motion fidelity.

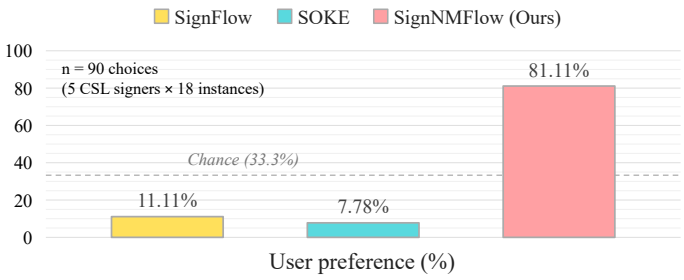

Figure 4: User study on the CSL-Daily: Ours vs. SOTA methods.

5 Conclusions

In this paper, we revisit SLP from the perspective of manifold transport and propose SignNMFlow, a naturalness-guided manifold flow matching framework. Unlike existing flow matching-based SLP methods that construct conditional paths by linear interpolation, SignNMFlow constructs conditional paths directly on the motion manifold, jointly considering geometric efficiency and the motion distribution. Specifically, we represent sign motions on the product of unit-quaternion hyperspheres and construct the conditional paths through the naturalness-guided interpolation, which couples a closed-form geodesic with a learnable deviation. To incorporate the motion distribution, we introduce a motion naturalness measure built from prototypical motions and learn the interpolation by minimizing the kinetic energy under this measure. Additionally, we adopt a target-prediction parameterization that provides a well-scaled supervision signal without extra losses. Extensive evaluations on Phoenix-2014T, CSL-Daily, and How2Sign validate the effectiveness of SignNMFlow.

AI USE STATEMENT

In this work, we have not used generative AI tools for any task that requires disclosure, including generating synthetic data, developing theoretical models or conceptual frameworks, formulating or proving mathematical claims, proposing hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, or interpreting results. We used generative AI tools only to edit the manuscript and improve the readability of the English text. We have reviewed all AI-assisted text and verified that the technical content, claims, and reported numbers originate from our own experiments and derivations. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies sign language production, whose long-term goal is to improve accessibility for Deaf and hard-of-hearing communities; the generated sign motions are intended to support, not replace, human interpreters. All experiments rely on three publicly released research benchmarks, Phoenix-2014T, CSL-Daily, and How2Sign, which are used for their intended research purpose, and no new signer data are collected for training. The user study involved five CSL signers who participated voluntarily with informed consent; the study records only aggregate selection counts over anonymized generated motions, and no personally identifiable information is collected or reported.

REPRODUCIBILITY STATEMENT

All experiments are conducted on three public benchmarks, Phoenix-2014T, CSL-Daily, and How2Sign; the datasets, evaluation metrics, and evaluation protocol are described in Sec. 4.1. The training and sampling procedures of SignNMFlow are summarized in Algorithm 1 and Algorithm 2. Complete architecture and training hyperparameters, including the configuration of the motion naturalness measure, are listed in Appendix D, and the full derivation of the geometric mismatch (Property 1) is provided in Appendix B. The comparison with prior work follows the protocol of SOKE (Zuo et al., 2025). We will release the source code upon publication.

REFERENCES

Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

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. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1985–1995, 2024.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7784–7793, 2018.

Ricky TQ Chen and Yaron Lipman. Flow matching on general geometries. In *International Conference on Learning Representations*, pp. 47922–47945, 2024.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. Two-stream network for sign language recognition and translation. *Neural Information Processing Systems*, pp. 17043–17056, 2022.

Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, pp. 5–30, 2006.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2sign: A large-scale multimodal dataset for continuous american sign language. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2734–2743, 2021.

Olaf Dünkel, Tim Salzmann, and Florian Pfaff. Normalizing flows on the product space of s0(3) manifolds for probabilistic human pose modeling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2285–2294, 2024.

John RW Glauert, Ralph Elliott, Stephen J Cox, Judy Tryggvason, and Mary Sheard. Vanessa–a system for communication between deaf and hearing people. *Technology and Disability*, pp. 207–216, 2006.

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. Llms are good sign language translators. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 18362–18372, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Neural Information Processing Systems*, pp. 6840–6851, 2020.

Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Continuous sign language recognition with correlation network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2529–2539, 2023a.

Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Self-emphasizing network for continuous sign language recognition. In *AAAI Conference on Artificial Intelligence*, pp. 854–862, 2023b.

Youngjoon Jang, Haran Raajesh, Liliane Momeni, Gül Varol, and Andrew Zisserman. Lost in translation, found in context: Sign language translation with contextual cues. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8742–8752, 2025.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Neural Information Processing Systems*, pp. 20067–20079, 2023.

Kostas Karpouzis, George Caridakis, S-E Fotinea, and Eleni Efthimiou. Educational resources and implementation of a greek sign language synthesis architecture. *Computers & Education*, pp. 54–74, 2007.

Nabeela Khan, 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 *International Conference on Multimodal Interaction*, pp. 173–180, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

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

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, pp. 726–742, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jianhe Low, Alexandre Symeonidis-Herzig, Maksym Ivashechkin, Ozge Mercanoglu Sincan, and Richard Bowden. Signspark: Efficient multilingual sign language production via sparse keyframe learning. In *European Conference on Computer Vision*, pp. 648–670, 2026.

James B McQueen. Some methods of classification and analysis of multivariate observations. In *Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10975–10985, 2019.

Bruno Pelletier. Kernel density estimation on riemannian manifolds. *Statistics & probability letters*, pp. 297–304, 2005.

David Pollard. Strong consistency of k-means clustering. *The annals of statistics*, pp. 135–140, 1981.

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, 2020.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265, 2015.

Stephanie Stoll, Armin Mustafa, and Jean-Yves Guillemaut. There and back again: 3d sign language generation from text using back-translation. In *International Conference on 3D Vision*, pp. 187–196, 2022.

JungHoon Sung, Boeun Kim, Chu Xin, Hyung Jin Chang, ChangHo Kim, Sang-Il Choi, and Younggeun Choi. Context-aware autoregressive diffusion for gloss-wise sign language production. *arXiv preprint arXiv:2606.21234*, 2026.

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

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. Sign2gpt: Leveraging large language models for gloss-free sign language translation. In *International Conference on Learning Representations*, volume 2024, pp. 18157–18174, 2024.

Aoxiong Yin, Haoyuan Li, Kai Shen, Siliang Tang, and Yueting Zhuang. T2s-gpt: Dynamic vector quantization for autoregressive sign language production from text. In *Association for Computational Linguistics*, pp. 3345–3356, 2024.

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*, pp. 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 20814–20824, 2023.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1316–1325, 2021.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5738–5746, 2019.

Ronglai Zuo, Fangyun Wei, and Brian Mak. Natural language-assisted sign language recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 14890–14900, 2023.

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

Appendix

Table of Contents

A Related Work **14**

A.1 Sign Language Production 14

A.2 Flow Matching 14

B Detailed Derivation of Property 1 **14**

B.1 Derivation on $SO(3)$ 14

B.2 Derivation on the 6D Rotation Representation 15

B.3 Derivation on the Quaternion Representation 16

B.4 Derivation on the Euler-Angle Representation 17

C Additional Details for Section 3 **18**

C.1 Quaternion Sign Convention 18

C.2 Prototype Extraction 19

C.3 Validity of the Motion Naturalness Measure 19

C.4 Exactness of the Naturalness-Guided Interpolation 22

C.5 Training and Sampling the Generation Network 23

D Supplementary Experiments **24**

D.1 Experimental Detail 24

D.2 Additional Experimental Results 25

A Related Work

A.1 Sign Language Production

Early SLP methods (Glauert et al., 2006; Karpouzis et al., 2007) primarily relied on rule-based systems and manually designed animation libraries, which achieved text-to-sign conversion by establishing mappings between textual semantics and predefined sign motions. However, due to their dependence on manually crafted rules and limited motion templates, these methods struggled to model the complex semantic relationships and continuous motion dynamics. Recent SLP works have mainly followed three families: autoregressive models, diffusion models, and flow-based generative models. Autoregressive methods (Saunders et al., 2020) generate sign motions sequentially, and their recent tokenizer-LM variants (Yin et al., 2024; Zuo et al., 2025) encode continuous sign motion into discrete latent representations to leverage language models for sequential dependencies. Diffusion methods (Baltatzis et al., 2024; Tang et al., 2025; Sung et al., 2026) generate sign motions by iterative denoising under text-conditioned guidance. Among them, GARD (Sung et al., 2026) refines inter-gloss transitions with the geodesic distance in place of the Euclidean one, yet the improvement remains local to gloss transitions and its generative transport stays in Euclidean space. Flow matching-based methods (Khan et al., 2025; Low et al., 2026) anchor training on conditional paths but construct these paths by linear interpolation, thereby ignoring the rotational geometry of human joints and the motion distribution of sign motions. These efforts thus center on conditioning and decoding mechanisms, while the geometry of the generative transport remains underexplored. In this work, we construct conditional paths on the motion manifold, jointly considering geometric efficiency and the motion distribution.

A.2 Flow Matching

Training continuous normalizing flows originally required simulating the ODE and backpropagating through the solver, which is computationally expensive. Flow matching (Lipman et al., 2022) and the concurrent stochastic-interpolant formulation (Albergo & Vanden-Eijnden, 2022) remove this cost by regressing a learnable velocity field onto per-pair conditional paths, yielding a simulation-free objective anchored on conditional paths. Subsequent work studies how these paths should be designed. Rectified flow (Liu et al., 2022) transports samples along the linear interpolation between paired endpoints and iteratively straightens the trajectories to enable few-step sampling, and mini-batch optimal transport (Tong et al., 2023) couples the endpoints to reduce path crossing. Flow matching on general geometries (Chen & Lipman, 2024) further replaces the linear interpolation with geodesic interpolation on Riemannian manifolds, so that conditional paths respect the intrinsic geometry of the data. Flow matching has also entered SLP. SignFlow (Khan et al., 2025) introduces conditional flow matching with optimal-transport coupling for text-driven sign motion generation, and SignSparK (Low et al., 2026) incorporates sparse keyframes into the flow matching framework to guide generation. Both methods construct their conditional paths by linear interpolation in Euclidean space, treating rotation representations as unconstrained Euclidean variables.

B Detailed Derivation of Property 1

In this section, we provide the detailed derivation that completes the proof of Property 1, and then extend the analysis to the remaining standard rotation representations. We first work directly on the rotation group $SO(3)$ (Sec. B.1), then show that the 6D rotation representation adopted by existing flow matching-based SLP methods inherits exactly the same off-manifold deviation (Sec. B.2), and finally establish the deviation for the quaternion representation (Sec. B.3) and the Euler-angle representation (Sec. B.4).

B.1 Derivation on $SO(3)$

Setup. Recall that $SO(3) = \{Q \in \mathbb{R}^{3\times 3} \mid Q^\top Q = I,\ \det(Q) = 1\}$, and every $R \in SO(3)$ has orthonormal columns and hence satisfies $\|R\|_F^2 = \mathrm{tr}(R^\top R) = 3$. Given two distinct rotations $R_0, R_1 \in SO(3)$, the conditional path constructed by linear interpolation (Eq. 1 in the main text) is

$$R_t = (1-t)\,R_0 + t\,R_1, \quad t \in [0,1]. \tag{14}$$

Derivation. Expanding $\|R_t\|_F^2 = \langle R_t, R_t\rangle_F$ using the bilinearity of the Frobenius inner product yields

$$\begin{aligned}\|R_t\|_F^2 &= (1-t)^2\,\|R_0\|_F^2 + 2t(1-t)\,\langle R_0, R_1\rangle_F + t^2\,\|R_1\|_F^2 \\ &= 3\big[(1-t)^2 + t^2\big] + 2t(1-t)\,\langle R_0, R_1\rangle_F. \end{aligned} \tag{15}$$

On the other hand, the squared Frobenius distance between the two endpoints expands as

$$\begin{aligned}\|R_0 - R_1\|_F^2 &= \|R_0\|_F^2 - 2\,\langle R_0, R_1\rangle_F + \|R_1\|_F^2 \\ &= 6 - 2\,\langle R_0, R_1\rangle_F, \end{aligned} \tag{16}$$

which gives $\langle R_0, R_1\rangle_F = 3 - \frac{1}{2}\|R_0 - R_1\|_F^2$. Substituting this relation into Eq. 15 and using the identity $(1-t)^2 + t^2 = 1 - 2t(1-t)$ completes the derivation:

$$\|R_t\|_F^2 = 3 - t(1-t)\,\|R_0 - R_1\|_F^2. \tag{17}$$

Moreover, writing θ for the relative rotation angle of $R_0^\top R_1$, the Frobenius inner product between two rotations admits the closed form $\langle R_0, R_1\rangle_F = \mathrm{tr}(R_0^\top R_1) = 1 + 2\cos\theta$, so that $\|R_0 - R_1\|_F^2 = 4(1-\cos\theta)$ and Eq. 17 reads $\|R_t\|_F^2 = 3 - 4t(1-t)(1-\cos\theta) < 3$ whenever $\theta \neq 0$, i.e., $R_0 \neq R_1$. The determinant shrinks in parallel: the eigenvalues of $R_0^\top R_1$ are 1 and $e^{\pm i\theta}$, so $\det(R_t) = \det\big(R_0\big[(1-t)I + t\,R_0^\top R_1\big]\big) = 1 - 2t(1-t)(1-\cos\theta) < 1$, and the orthogonality and determinant constraints of $SO(3)$ are violated simultaneously.

Completion of Property 1. If $R_0 \neq R_1$, then $\|R_0 - R_1\|_F^2 > 0$. For any $t \in (0,1)$ we have $t(1-t) > 0$, and Eq. 17 yields

$$\|R_t\|_F^2 = 3 - t(1-t)\,\|R_0 - R_1\|_F^2 < 3. \tag{18}$$

Since every $R \in SO(3)$ satisfies $\|R\|_F^2 = 3$, the interpolated state R_t violates the orthogonality constraint $R_t^\top R_t = I$, and therefore $R_t \notin SO(3)$ for all $t \in (0,1)$.

B.2 Derivation on the 6D Rotation Representation

In practice, existing flow matching-based SLP methods (Khan et al., 2025; Low et al., 2026) represent joint rotations with the 6D rotation representation (Zhou et al., 2019) $\psi(R) = (a, b)$, which retains the first two columns of R. Since ψ is a linear map, $\psi(R_t) = (1-t)\,\psi(R_0) + t\,\psi(R_1)$, so the off-manifold deviation established in Sec. B.1 persists on the embedded 6D manifold. We make this precise at the coordinate level below.

Setup. The manifold of valid 6D representations is $\mathcal{M}_{6\mathrm{D}} = \{(a,b) \in \mathbb{R}^3 \times \mathbb{R}^3 \mid \|a\|_2 = \|b\|_2 = 1,\ a^\top b = 0\}$. Given two points $x_0 = (a_0, b_0)$ and $x_1 = (a_1, b_1)$ on $\mathcal{M}_{6\mathrm{D}}$, the conditional path constructed by linear interpolation is

$$x_t = (1-t)\,x_0 + t\,x_1 = (a_t,\ b_t), \quad t \in [0,1], \tag{19}$$

where $a_t = (1-t)a_0 + ta_1$ and $b_t = (1-t)b_0 + tb_1$. Since $x_0, x_1 \in \mathcal{M}_{6\mathrm{D}}$, their components satisfy the unit-norm constraints

$$\|a_0\|_2^2 = \|a_1\|_2^2 = \|b_0\|_2^2 = \|b_1\|_2^2 = 1. \tag{20}$$

Derivation. Expanding the squared norm of a_t via the bilinearity of the inner product yields

$$\begin{aligned}\|a_t\|_2^2 &= \big\|(1-t)a_0 + ta_1\big\|_2^2 \\ &= \big\langle (1-t)a_0 + ta_1,\ (1-t)a_0 + ta_1\big\rangle \\ &= (1-t)^2\,\|a_0\|_2^2 + 2t(1-t)\,a_0^\top a_1 + t^2\,\|a_1\|_2^2. \end{aligned} \tag{21}$$

Substituting the unit-norm constraints $\|a_0\|_2^2 = \|a_1\|_2^2 = 1$ into Eq. 21 and using the identity $(1-t)^2 + t^2 = 1 - 2t(1-t)$, we obtain

$$\begin{aligned}\|a_t\|_2^2 &= (1-t)^2 + t^2 + 2t(1-t)\,a_0^\top a_1 \\ &= 1 - 2t(1-t)\big(1 - a_0^\top a_1\big). \end{aligned} \tag{22}$$

On the other hand, the squared distance between the two endpoints expands as

$$\begin{aligned}\|a_0 - a_1\|_2^2 &= \|a_0\|_2^2 - 2\, a_0^\top a_1 + \|a_1\|_2^2 \\ &= 2\big(1 - a_0^\top a_1\big),\end{aligned} \tag{23}$$

which gives $1 - a_0^\top a_1 = \frac{1}{2}\|a_0 - a_1\|_2^2$. Substituting this relation into Eq. 22 completes the derivation:

$$\|a_t\|_2^2 = \big\|(1-t)a_0 + t a_1\big\|_2^2 = 1 - t(1-t)\,\|a_0 - a_1\|_2^2. \tag{24}$$

Completion. If $a_0 \neq a_1$, then by the Cauchy–Schwarz inequality $a_0^\top a_1 \leq \|a_0\|_2 \|a_1\|_2 = 1$, with equality if and only if $a_0 = a_1$; hence $\|a_0 - a_1\|_2^2 = 2(1 - a_0^\top a_1) > 0$ by Eq. 23. For any $t \in (0, 1)$ we have $t(1-t) > 0$, and Eq. 24 yields

$$\|a_t\|_2^2 = 1 - t(1-t)\,\|a_0 - a_1\|_2^2 < 1, \tag{25}$$

so a_t violates the unit-norm constraint. If instead $a_0 = a_1$ but $x_0 \neq x_1$, then $b_0 \neq b_1$, and the same argument applied to the b component shows $\|b_t\|_2^2 < 1$. In either case, at least one component of x_t violates the defining constraints of $\mathcal{M}_{6\mathrm{D}}$, and therefore $x_t \notin \mathcal{M}_{6\mathrm{D}}$ for all $t \in (0, 1)$, in exact correspondence with the matrix-level deviation established in Sec. B.1.

B.3 Derivation on the Quaternion Representation

Unit quaternions are the most compact minimal parameterization of rotations: each rotation corresponds to a point of $\mathbb{S}^3 = \{q \in \mathbb{R}^4 \mid \|q\|_2 = 1\}$, and $\mathbb{S}^3$ is the representation adopted by SignNM-Flow (Sec. 3.1). We show that the conditional path constructed by linear interpolation between unit quaternions leaves $\mathbb{S}^3$ through exactly the same norm-contraction mechanism as on $SO(3)$ and the 6D representation.

Setup. Given two distinct unit quaternions $q_0, q_1 \in \mathbb{S}^3$, the conditional path constructed by linear interpolation is

$$q_t = (1-t)\, q_0 + t\, q_1, \qquad t \in [0, 1], \tag{26}$$

where both endpoints satisfy the unit-norm constraint $\|q_0\|_2^2 = \|q_1\|_2^2 = 1$.

Derivation. Expanding the squared norm of q_t via the bilinearity of the inner product yields

$$\begin{aligned}\|q_t\|_2^2 &= (1-t)^2\,\|q_0\|_2^2 + 2t(1-t)\, q_0^\top q_1 + t^2\,\|q_1\|_2^2 \\ &= 1 - 2t(1-t)\big(1 - q_0^\top q_1\big).\end{aligned} \tag{27}$$

On the other hand, the squared distance between the two endpoints expands as

$$\begin{aligned}\|q_0 - q_1\|_2^2 &= \|q_0\|_2^2 - 2\, q_0^\top q_1 + \|q_1\|_2^2 \\ &= 2\big(1 - q_0^\top q_1\big),\end{aligned} \tag{28}$$

which gives $1 - q_0^\top q_1 = \frac{1}{2}\|q_0 - q_1\|_2^2$. Substituting this relation into Eq. 27 completes the derivation:

$$\|q_t\|_2^2 = \big\|(1-t)\, q_0 + t\, q_1\big\|_2^2 = 1 - t(1-t)\,\|q_0 - q_1\|_2^2. \tag{29}$$

Writing $\theta = \arccos(q_0^\top q_1) \in [0, \pi]$ for the spherical distance between the two representatives, Eq. 28 gives $\|q_0 - q_1\|_2^2 = 2(1 - \cos\theta) = 4\sin^2(\theta/2)$, so the deviation also reads $\|q_t\|_2^2 = 1 - 4t(1-t)\sin^2(\theta/2)$, in exact correspondence with the angle form $\|R_t\|_F^2 = 3 - 4t(1-t)(1-\cos\theta)$ of Sec. B.1, with θ playing the role of the half rotation angle.

Completion. If $q_0 \neq q_1$, then by the Cauchy–Schwarz inequality $q_0^\top q_1 \leq \|q_0\|_2 \|q_1\|_2 = 1$, with equality if and only if $q_0 = q_1$; hence $\|q_0 - q_1\|_2^2 > 0$ by Eq. 28. For any $t \in (0, 1)$ we have $t(1-t) > 0$, and Eq. 29 yields

$$\|q_t\|_2^2 = 1 - t(1-t)\,\|q_0 - q_1\|_2^2 < 1, \tag{30}$$

so q_t violates the unit-norm constraint that defines $\mathbb{S}^3$, and therefore $q_t \notin \mathbb{S}^3$ for all $t \in (0, 1)$. On the motion manifold $\mathcal{M} = (\mathbb{S}^3)^J$ of Sec. 3.1, linear interpolation between two motions applies this computation joint by joint, so every joint of every interior state leaves its hypersphere factor. The deviation is thus present in the raw quaternion coordinates as well, prior to any manifold structure

being imposed, and it is again the Exp-based construction of Sec. 3.3 (Appendix C.4) that restores manifold validity.

Remark (renormalized interpolation). A natural remedy is to project the interpolated state back onto the sphere, $\hat{q}_t = q_t/\|q_t\|_2$, which yields normalized linear interpolation (nlerp). In the two-plane $\mathrm{span}\{q_0, q_1\}$, write $q_0 = e_1$ and $q_1 = \cos\theta\, e_1 + \sin\theta\, e_2$ with $\theta \in (0, \pi)$; the renormalized state $\hat{q}_t$ traces the same great-circle arc as the spherical interpolation (slerp), but at the state-dependent arc angle $\chi(t) = \mathrm{atan2}\big(t\sin\theta,\ 1 - t + t\cos\theta\big)$, and differentiating gives $\dot{\chi}(t) = \sin\theta/\|q_t\|_2^2$. Since $\|q_t\|_2^2$ attains its minimum at $t = \frac{1}{2}$, the angular speed peaks mid-path and exceeds its endpoint value by the factor

$$\frac{\dot{\chi}(\frac{1}{2})}{\dot{\chi}(0)} = \frac{1}{\|q_{1/2}\|_2^2} = \frac{2}{1+\cos\theta} = \sec^2\big(\tfrac{\theta}{2}\big), \tag{31}$$

so the renormalized path departs from the constant-speed geodesic whenever$\theta \neq 0$, and the distortion grows without bound as the representatives approach antipodes ($\theta \to \pi$). Renormalization therefore restores the constraint but not geometric efficiency: the states sweep the connecting arc at a state-dependent speed, in contrast to the constant-speed geodesic reference that Sec. 3.3 adopts.

B.4 Derivation on the Euler-Angle Representation

Euler angles parameterize a rotation by three angles $u = (\alpha, \beta, \gamma)$ through a fixed axis factorization. We take the ZYX convention $R(u) = R_z(\alpha)\, R_y(\beta)\, R_x(\gamma)$, where R_z, R_y, R_x denote rotations about the axes of the standard Euclidean frame (the other conventions are analogous). Unlike the representations above, the raw parameter space $\mathbb{R}^3$ carries no algebraic constraint: every angle triple parameterizes a valid rotation. The incompatibility between linear interpolation and rotational geometry therefore manifests through two complementary mechanisms, which we derive in turn: violating the periodic structure of the faithful parameter manifold, and distorting the induced rotation path when the raw angles are treated as unconstrained Euclidean variables.

Derivation I: periodicity mismatch on the faithful parameter manifold. Each Euler angle is defined modulo 2π, so the faithful parameter manifold is the torus $\mathbb{T}^3 = (\mathbb{S}^1)^3$, embedded in $\mathbb{R}^6$ through the standard circle embedding $\eta(\alpha) = (\cos\alpha, \sin\alpha) \in \mathbb{S}^1 \subset \mathbb{R}^2$ applied per angle. Linear interpolation between two embedded representations $x_0, x_1 \in \mathbb{T}^3$ is the per-coordinate interpolation $x_t = (1-t)x_0 + t\,x_1$ in $\mathbb{R}^6$. Fixing one angle and writing $\Delta = \alpha_1 - \alpha_0$ for its increment, the bilinearity of the inner product gives

$$\begin{aligned}\big\|(1-t)\,\eta(\alpha_0) + t\,\eta(\alpha_1)\big\|_2^2 &= (1-t)^2 + t^2 + 2t(1-t)\cos\Delta \\ &= 1 - 2t(1-t)\big(1-\cos\Delta\big) \\ &= 1 - t(1-t)\,\big\|\eta(\alpha_0) - \eta(\alpha_1)\big\|_2^2,\end{aligned} \tag{32}$$

using $\|\eta(\alpha_0) - \eta(\alpha_1)\|_2^2 = 2 - 2\cos\Delta = 4\sin^2(\Delta/2) > 0$ whenever $\alpha_0 \not\equiv \alpha_1 \pmod{2\pi}$. Each circle factor of the interpolated state therefore violates its unit-norm constraint for every $t \in (0,1)$, with the deviation $4t(1-t)\sin^2(\Delta/2)$ maximized, over the increment, when the two angles differ by an odd multiple of π. This is the norm-contraction of Secs. B.1–B.3 instantiated on $\mathbb{S}^1$. The conditional path constructed by linear interpolation leaves the embedded torus $\mathbb{T}^3$ whenever the two representations differ.

Derivation II: distortion of raw-angle interpolation. In practice, Euler angles would be stored and interpolated as unconstrained Euclidean vectors, $u_t = (1-t)\,u_0 + t\,u_1 \in \mathbb{R}^3$, the Euler analogue of treating rotation coordinates as flat Euclidean variables (Sec. 3.1). Every intermediate triple is a valid parameterization, so no intermediate state is formally off-manifold; the failure instead appears in the geometry of the induced rotation path $t \mapsto R(u_t)$, in two quantifiable ways.

First, the transport is state-dependent. Let $\omega_b(t) \in \mathbb{R}^3$ denote the angular velocity of $R(u_t)$ expressed in the moving frame, i.e., the unique vector satisfying $\dot{R}(u_t) = R(u_t)\,[\omega_b(t)]_\times$, where $[\cdot]_\times$ is the cross-product matrix. Differentiating the factorization gives

$$\omega_b(t) = \dot{\gamma}\,\hat{x} + \dot{\beta}\,R_x(\gamma)^\top \hat{y} + \dot{\alpha}\,R_x(\gamma)^\top R_y(\beta)^\top \hat{z} = J(u_t)\,(u_1 - u_0), \tag{33}$$

with the constant rate $\dot{u}_t = u_1 - u_0$ and the state-dependent Jacobian

$$J(u) = \begin{pmatrix} -\sin\beta & 0 & 1 \\ \cos\beta\,\sin\gamma & \cos\gamma & 0 \\ \cos\beta\,\cos\gamma & -\sin\gamma & 0 \end{pmatrix}, \qquad \det J(u) = -\cos\beta, \tag{34}$$

whose determinant vanishes exactly on the gimbal-lock locus $\{\beta = \frac{\pi}{2} + k\pi,\ k \in \mathbb{Z}\}$. Along the geodesic, the rotation proceeds at constant speed about a fixed axis; along the Euler-interpolated path, both the direction and the magnitude of $\omega_b(t)$ vary with t through u_t. At gimbal lock, J loses rank and the parameterization fails to be a local diffeomorphism: since $\sigma_{\min}\big(J(u)\big) \le |\det J(u)|^{1/3} \to 0$ as $\beta \to \frac{\pi}{2} + k\pi$ (while $\|J(u)\|_F = \sqrt{3}$ for every u, so the collapse is purely anisotropic), the inverse map degenerates, $\|J(u)^{-1}\|_2 = 1/\sigma_{\min}\big(J(u)\big) \to \infty$. Bounded Euclidean increments of the angles thus map to collapsing rotation increments in the near-kernel direction, while the angle rates required to realize angular velocities in the corresponding output direction diverge. Euclidean increments in parameter space carry no consistent rotation-geometric meaning near this locus.

Second, the parameterization is non-injective, and Euclidean interpolation fabricates spurious excursions between representations of the same rotation. Consider $u_0 = (0, 0, 0)$ and $u_1 = (\pi, \pi, \pi)$: the half-turn rotations $R_z(\pi)$, $R_y(\pi)$, $R_x(\pi)$ are diagonal with diagonal entries $(-1, -1, 1)$, $(-1, 1, -1)$, $(1, -1, -1)$, whose product is the identity, so both triples parameterize $R = I$ and the geodesic transport between them is the constant path, with geodesic distance $d\big(R(u_0), R(u_1)\big) = 0$. The linearly interpolated path $u_t = t\,(\pi, \pi, \pi)$ is nevertheless non-constant, and its midpoint evaluates to

$$R\big(\tfrac{\pi}{2}, \tfrac{\pi}{2}, \tfrac{\pi}{2}\big) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{pmatrix} = R_y\big(\tfrac{\pi}{2}\big), \tag{35}$$

a rotation of angle $\frac{\pi}{2}$ about the y-axis. Euclidean interpolation in angle space therefore produces an intermediate state at angular distance at least $\frac{\pi}{2}$ from both endpoints,

$$\sup_{t\in[0,1]}\ d\big(R(u_t),\, R(u_0)\big) \ \ge\ \tfrac{\pi}{2} \ >\ 0 \ =\ d\big(R(u_0),\, R(u_1)\big), \tag{36}$$

where $d(\cdot, \cdot)$ denotes the geodesic distance on $SO(3)$ (the rotation angle of the relative rotation). The spurious excursion is an instance of the standard aliasing identity $R(\alpha{+}\pi,\ \pi{-}\beta,\ \gamma{+}\pi) = R(\alpha, \beta, \gamma)$, under which u_1 is an alternative parameterization of u_0: because the parameterization folds $\mathbb{R}^3$ onto $SO(3)$, linear structure in parameter space carries no rotation-geometric meaning.

Completion. For the Euler-angle representation, the faithful parameter manifold $\mathbb{T}^3$ suffers the same off-manifold deviation as $SO(3)$ (Sec. B.1), the 6D representation (Sec. B.2), and the quaternion representation (Sec. B.3) (Derivation I), while interpolating the constraint-free raw angles substitutes a complementary failure: a state-dependent, rank-deficient, and non-injective distortion of the induced rotation path (Derivation II). Both mechanisms stem from applying Euclidean-linear operations to rotation parameters. Across the four standard rotation representations, linear interpolation therefore either leaves the representation manifold or distorts the rotation path itself; constructing conditional paths on $\mathcal{M}$ through the exponential map (Sec. 3.3) avoids both failure modes.

C Additional Details for Section 3

C.1 Quaternion Sign Convention

Unit quaternions q and $-q$ represent the same rotation. Under the scalar-first convention (q_w, q_x, q_y, q_z), we resolve this sign ambiguity by selecting the representative whose first nonzero component is positive. For each quaternion $\tilde{q}_j = (\tilde{q}_{j,w}, \tilde{q}_{j,x}, \tilde{q}_{j,y}, \tilde{q}_{j,z})$ of the motion x in Eq. 5, the sign selection is

$$q_j = \begin{cases} -\tilde{q}_j, & \tilde{q}_{j,w} < 0, \\ \tilde{q}_j, & \tilde{q}_{j,w} > 0, \\ \sigma_j \tilde{q}_j, & \tilde{q}_{j,w} = 0, \end{cases} \tag{37}$$

where $\tilde{q}_j$ is the quaternion before sign adjustment, q_j is the selected representative, and $\sigma_j \in \{+1, -1\}$ makes the first nonzero component of $(\tilde{q}_{j,x}, \tilde{q}_{j,y}, \tilde{q}_{j,z})$ positive. This selects a unique

representative for every rotation, including the 180° case. All representatives lie in a closed hemisphere of $\mathbb{S}^3$, and no two of them are antipodal, so the logarithmic maps in Eq. 6 are always evaluated between non-antipodal quaternions.

C.2 Prototype Extraction

Given a training set containing N motion frames, we extract prototypical motions to characterize the motion distribution. For each joint independently, we perform K-means clustering (McQueen, 1967) over all joint rotations and obtain K prototypes. The process can be formalized as:

$$\{\tilde{c}_k^j\}_{k=1}^K = \arg\min_{\{\mu_k\}_{k=1}^K} \sum_{n=1}^N \min_{k\in\{1,\dots,K\}} \left\|q_n^j - \mu_k\right\|_2^2, \quad c_k^j = \frac{\tilde{c}_k^j}{\|\tilde{c}_k^j\|_2}, \quad k = 1,\dots,K, \tag{38}$$

where q_n^j denotes the sign-aligned quaternion in Eq. 37 of joint j at frame n, and c_k^j is the normalized prototype. To preserve the occurrence information of the motions, we further associate each prototype with an empirical weight:

$$w_k^j = \frac{|\mathcal{I}_k^j|}{N}, \qquad \sum_{k=1}^K w_k^j = 1, \tag{39}$$

where $\mathcal{I}_k^j$ denotes the set of rotations assigned to the k-th cluster of the j-th joint, and w_k^j represents the corresponding empirical occurrence probability. The prototypes $\{c_k\}_{k=1}^K$ and frequency weights $\{w_k\}_{k=1}^K$ in Sec. 3.2 are instantiated joint-wise as $c_k = c_k^j$ and $w_k = w_k^j$ within each factor of the product manifold. The renormalized centroid c_k^j coincides with the spherical K-means centroid and closely approximates the Riemannian (Fréchet) mean whenever the cluster is concentrated.

C.3 Validity of the Motion Naturalness Measure

We establish the validity of the measure through four analyses: G is well defined on $\mathcal{M}$, the density factor consistently estimates the motion distribution, the induced transport cost is inversely weighted by motion density, and the measure degenerates to the standard geometry under uniform distribution.

Setup and notation. The naturalness measure reads $G(x) = m(x)\big(S(x) + \rho\mathbf{I}\big)^{-1}$, where $m(x)$ denotes the aggregate kernel mass. Recall that $\delta_k = \mathrm{Log}_x(c_k) \in T_x\mathcal{M}$ with $\|\delta_k\| = d_{\mathcal{M}}(x, c_k)$, the geodesic distance on $\mathcal{M}$, and that $T_x\mathcal{M} \cong \mathbb{R}^{3J}$ carries the Euclidean inner product. Each prototype retains a non-empty cluster, so its frequency weight satisfies $w_k > 0$.

Well-definedness. The naturalness norm (Eq. 9) and the kinetic energy objective (Eq. 12) require $G(x)$ to be invertible at every point of the manifold, and the regularized construction in Eq. 8 guarantees this. Since $w_k > 0$ and the Gaussian kernel is strictly positive, $\tilde{w}_k(x) > 0$ for all k and all x, hence $m(x) > 0$. Each outer product $\delta_k\delta_k^\top$ is symmetric positive semidefinite (PSD), and a nonnegative combination of PSD matrices is PSD, so $S(x) \succeq 0$ and therefore $S(x) + \rho\mathbf{I} \succeq \rho\mathbf{I} \succ 0$. The inverse of a symmetric positive definite (SPD) matrix is SPD, and a positive multiple of an SPD matrix is SPD, so $G(x) = m(x)\big(S(x) + \rho\mathbf{I}\big)^{-1} \succ 0$ at every $x \in \mathcal{M}$. Consequently, $G(x)^{-1}$ exists on $\mathcal{M}$, and $\mathcal{N}(\gamma)$ (Eq. 9), $\varphi(\gamma)$ (Eq. 10), and $\mathcal{L}_G(\theta)$ (Eq. 12) are well-defined. Moreover, $G(x)^{-1} = \frac{1}{m(x)}\big(S(x) + \rho\mathbf{I}\big)$ has eigenvalues in $\big[\rho/m(x),\ (\lambda_{\max}(S(x)) + \rho)/m(x)\big]$, and the Rayleigh quotient of a symmetric matrix lies between its extreme eigenvalues, so for every $v \in T_x\mathcal{M}$,

$$\frac{\rho}{m(x)}\,\|v\|^2 \ \le\ v^\top G(x)^{-1} v \ \le\ \frac{\lambda_{\max}\big(S(x)\big) + \rho}{m(x)}\,\|v\|^2, \tag{40}$$

where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue. These bounds quantify how the regularization constant ρ and the aggregate density $m(x)$ jointly control the transport cost.

Density consistency. We show that the density factor $m(x)$ is a kernel density estimate (KDE) of the motion distribution on $\mathcal{M}$, compressed onto the K prototypes with error controlled by K. Let $\{x_i\}_{i=1}^N$ be the training motions. Let $\mathcal{C}_k$ be the cluster assigned to prototype c_k, with frequency weight $w_k = |\mathcal{C}_k|/N$, and let $\kappa_\sigma(r) = \exp(-r^2/2\sigma^2)$. Eq. 8 then gives

$m(x) = \frac{1}{N}\sum_{k=1}^{K} |\mathcal{C}_k|\, \kappa_\sigma\big(d_{\mathcal{M}}(x, c_k)\big)$. Let $\hat{p}_\sigma(x) = \frac{1}{N}\sum_{i=1}^{N} \kappa_\sigma\big(d_{\mathcal{M}}(x, x_i)\big)$ be the KDE built directly from the training motions. The geodesic distance satisfies the triangle inequality, so $\big|d_{\mathcal{M}}(x, x_i) - d_{\mathcal{M}}(x, c_{k(i)})\big| \le d_{\mathcal{M}}(x_i, c_{k(i)})$, where $k(i)$ indexes the cluster containing x_i. The kernel κ_σ is globally Lipschitz with constant $L_\sigma = \max_{r\ge 0} |\kappa'_\sigma(r)| = \frac{1}{\sigma\sqrt{e}}$, attained at $r = \sigma$, so each summand obeys $\big|\kappa_\sigma(d_{\mathcal{M}}(x, x_i)) - \kappa_\sigma(d_{\mathcal{M}}(x, c_{k(i)}))\big| \le \frac{1}{\sigma\sqrt{e}}\, d_{\mathcal{M}}(x_i, c_{k(i)})$, and averaging over i yields

$$\big|m(x) - \hat{p}_\sigma(x)\big| \;\le\; \frac{\bar{r}_K}{\sigma\sqrt{e}}, \qquad \bar{r}_K := \frac{1}{N}\sum_{i=1}^{N} d_{\mathcal{M}}\big(x_i,\, c_{k(i)}\big). \tag{41}$$

Moreover, the optimal quantization error $\bar{r}^\star_K$, the minimum of $\bar{r}_K$ over all K-prototype configurations, is non-increasing in K: any optimal K-prototype configuration induces a feasible $(K{+}1)$-prototype configuration of identical cost by duplicating one prototype, so $\bar{r}^\star_{K+1} \le \bar{r}^\star_K$. That $\bar{r}^\star_K \to 0$ as $K \to \infty$ for data on a compact manifold is a standard result in quantization theory (Pollard, 1981).

The kernel κ_σ depends on the pair (x, c) only through the geodesic distance, replacing it with the heat kernel profile, which on the homogeneous manifold $\mathcal{M}$ is likewise a fixed radial function of $d_{\mathcal{M}}$. Replacing κ_σ with this profile therefore turns $m(x)$ into a standard KDE on $\mathcal{M}$. Its consistency with the population motion density as $N \to \infty$ and $\sigma \to 0$ follows from the classical theory of kernel density estimation on Riemannian manifolds (Pelletier, 2005). The density factor of G is therefore a consistent estimate of the motion distribution. Compression onto prototypes adds the error $\bar{r}_K$, which is controlled by K (Eq. 41).

Distribution-aware transport cost. We next show how the measure injects the motion distribution into the transport cost. Expanding the quadratic form of $S(x)$ (Eq. 7) gives $v^\top S(x) v = \sum_{k=1}^{K} \tilde{w}_k(x)\, \langle \delta_k, v\rangle^2$, and substituting this into the identity $G(x)^{-1} = \frac{1}{m(x)}\big(S(x) + \rho\mathbf{I}\big)$ from the well-definedness analysis yields, for every $x \in \mathcal{M}$ and $v \in T_x\mathcal{M}$,

$$v^\top G(x)^{-1} v = \frac{1}{m(x)}\Big(\rho\,\|v\|^2 + \sum_{k=1}^{K} \tilde{w}_k(x)\, \big\langle \delta_k,\, v\big\rangle^2\Big), \tag{42}$$

whose numerator is a sum of nonnegative terms. The density estimate thus enters the instantaneous transport cost $\|\dot{\gamma}_t\|^2_{G(\gamma_t)}$ of Eq. 9 only through the prefactor $1/m(x)$: at a fixed numerator, transport through dense regions of the motion distribution contributes less to $\mathcal{N}(\gamma)$ and $\varphi(\gamma)$.

Equivalently, define the kernel-weighted covariance of the prototypes around x,

$$\Sigma(x) := \frac{S(x)}{m(x)} = \sum_{k=1}^{K} \frac{\tilde{w}_k(x)}{m(x)}\, \delta_k \delta_k^\top, \tag{43}$$

and factor $S(x) = m(x)\Sigma(x)$ out of Eq. 8:

$$G(x) = \Big(\Sigma(x) + \frac{\rho}{m(x)}\mathbf{I}\Big)^{-1}. \tag{44}$$

This form makes the semantics of naturalness explicit. The metric $G(x)$ is large near the core of a dense, tightly concentrated cluster (large $m(x)$, small $\Sigma(x)$) and small in sparse regions or locally ambiguous zones where the prototypes around x disperse. The transport cost in Eq. 42 is governed by $G(x)^{-1} = \Sigma(x) + \frac{\rho}{m(x)}\mathbf{I}$. It approaches the isotropic form $\frac{\rho}{m(x)}\|\cdot\|^2$ within tight cluster cores, while transit through sparse or dispersed regions is expensive. The construction parallels density-normalized diffusion geometry, which reweights the metric by the inverse data density so that transport preferentially follows the data (Coifman & Lafon, 2006). $\mathcal{L}_G$ in Eq. 12 therefore lowers the transport cost in regions where sign motions distribute, incorporating the motion distribution into the learned paths.

Geodesic recovery under a uniform distribution. Finally, we formalize the claim in Sec. 3.3 that the naturalness measure degenerates to the standard geometry when the motion distribution is uniform. The learned correction in Eq. 11 is then driven solely by the non-uniformity of the distribution. Let μ be the uniform probability measure on $\mathcal{M}$ (the normalized Riemannian volume),

and define the population counterparts of $m(x)$ (Eq. 8) and the scatter matrix (Eq. 7):

$$m_\mu(x) = \int_{\mathcal{M}} \kappa_\sigma\big(d_{\mathcal{M}}(x,c)\big)\, d\mu(c),$$
$$S_\mu(x) = \int_{\mathcal{M}} \kappa_\sigma\big(d_{\mathcal{M}}(x,c)\big)\, \mathrm{Log}_x(c)\, \mathrm{Log}_x(c)^\top\, d\mu(c), \tag{45}$$

where the integrals exclude the μ-null cut locus of x. In this regime, the population measure $G_\mu(x) := m_\mu(x)\big(S_\mu(x) + \rho\mathbf{I}\big)^{-1}$ is a constant multiple of the identity. $\varphi(\gamma)$ (Eq. 10) becomes positively proportional to the standard path energy $\int_0^1 \|\dot\gamma_t\|^2\, dt$. Its minimizer over absolutely continuous paths with $\gamma_0 = x_0$, $\gamma_1 = x_1$ is, up to reparameterization, the constant-speed geodesic $\gamma_t = \mathrm{Exp}_{x_0}\big(t\,\mathrm{Log}_{x_0}(x_1)\big)$, unique whenever x_1 does not lie in the cut locus of x_0.

To see that G_μ is constant and isotropic, note that the uniform measure μ is invariant under every isometry g of the round product manifold $\mathcal{M}$, isometries preserve geodesic distances, and they commute with the logarithmic map, $\mathrm{Log}_{gx}(gc) = g_*\,\mathrm{Log}_x(c)$. The change of variables $c \mapsto g^{-1}c$ in Eq. 45 therefore gives $m_\mu(gx) = m_\mu(x)$ and $S_\mu(gx) = g_*\, S_\mu(x)\, g_*^\top$. The isometry group of $\mathcal{M} = (\mathbb{S}^3)^J$ acts transitively, so $m_\mu \equiv m_0$ is constant. Taking g in the stabilizer of x, which acts on $T_x\mathcal{M} = \bigoplus_{j=1}^J T_{x^j}\mathbb{S}^3 \cong \mathbb{R}^{3J}$ as arbitrary independent rotations of the J three-dimensional factors together with factor permutations, the identity $S_\mu(x) = g_* S_\mu(x) g_*^\top$ forces $S_\mu(x) = \alpha\mathbf{I}$: blockwise rotation invariance makes $S_\mu(x)$ block-diagonal with scalar blocks, and permutation invariance equalizes the blocks. Transitivity then makes α independent of x, and $G_\mu \equiv \frac{m_0}{\alpha+\rho}\mathbf{I} =: c_g\mathbf{I}$.

To identify the energy minimizer, write the expectation over $t \sim \mathcal{U}[0,1]$ as an integral, so that $\varphi(\gamma) = c_g^{-1}\int_0^1 \|\dot\gamma_t\|^2\, dt$. For any absolutely continuous γ joining x_0 to x_1, the Cauchy–Schwarz inequality gives

$$\int_0^1 \|\dot\gamma_t\|^2\, dt \;\geq\; \Big(\int_0^1 \|\dot\gamma_t\|\, dt\Big)^2 = \ell(\gamma)^2 \;\geq\; d_{\mathcal{M}}(x_0,x_1)^2, \tag{46}$$

where $\ell(\gamma)$ is the path length and the last step is the definition of the geodesic distance as an infimum of lengths. Equality holds throughout if and only if γ travels at constant speed along a minimal-length path, i.e., γ is a minimizing geodesic. On $\mathcal{M}$, this geodesic is exactly $\gamma_t = \mathrm{Exp}_{x_0}(t\,\mathrm{Log}_{x_0}(x_1))$ traversed at constant speed. The constant-speed geodesic belongs to the parameterized family of Eq. 11 (attained at $\epsilon_\theta \equiv 0$). The minimum of $\mathcal{L}_G$ in Eq. 12 is therefore attained by a zero correction in this regime.

Remark (frame equivariance). For any isometry g of $\mathcal{M}$ applied simultaneously to the prototypes and the query point, $\tilde w_k(gx) = \tilde w_k(x)$ and $\delta_k(gx) = g_*\,\delta_k(x)$ with respect to the transformed prototypes, so $m(gx) = m(x)$ and $G(gx) = g_*\, G(x)\, g_*^{-1}$. The naturalness of a motion therefore depends only on the motion configuration, not on the choice of global reference frame.

Discussion. In the general (non-uniform) case, the exact minimizer of φ is the geodesic of the data-dependent Riemannian metric with quadratic form $v^\top G(x)^{-1} v$. Eq. 11 seeks this geodesic within a family anchored at the intrinsic geodesic, so geometric efficiency and naturalness are obtained on equal footing (Sec. 3.3). The four analyses above jointly certify the measure in Sec. 3.2. Well-definedness makes the optimization in Eq. 12 well-posed on the entire manifold. Density consistency grounds the measure in a consistent density estimate with compression error controlled by K, matching the underfitting side of the K-ablation in Table 4. The mild degradation at $K = 750$ instead reflects the variance of frequency weights in over-fragmented clusters, an effect the quantization bound does not model. The transport-cost analysis exposes the mechanism: naturalness reweights the ambient kinetic energy by the inverse motion density and the local dispersion of the prototypes (Eq. 44). Geodesic recovery shows that the learned correction is driven solely by the motion distribution while the geodesic anchor preserves geometric efficiency, so neither consideration is obtained at the expense of the other. It also explains the σ-ablation. A large σ flattens the kernel weights toward the uniform regime and weakens the distributional guidance of the measure, whereas a small σ amplifies prototype-level quantization noise (Eq. 41).

C.4 Exactness of the Naturalness-Guided Interpolation

In this subsection, we show that the naturalness-guided interpolation resides exactly on the motion manifold: with the projected correction of Eq. 48, the interpolation of Eq. 11 satisfies $x_{t,\theta} \in \mathcal{M}$ for all $t \in [0,1]$, for any raw network output and at any stage of training. The conditional path constructed by linear interpolation, in contrast, violates the defining constraint of $SO(3)$ at every interior time step (Property 1). The derivation uses only the closed forms of the exponential and logarithmic maps on $\mathbb{S}^3$ and the fact that both endpoints x_0 and x_1 are tuples of unit quaternions; no assumption is imposed on the raw correction $\tilde{\epsilon}_\theta$. We first collect the closed forms on the manifold (Sec. C.4.1), and then establish the exactness result (Sec. C.4.2).

C.4.1 Closed Forms on the Manifold

For the hypersphere $\mathbb{S}^3 = \{q \in \mathbb{R}^4 : \|q\| = 1\}$, the tangent space at $q \in \mathbb{S}^3$ is the linear subspace $T_q\mathbb{S}^3 = \{v \in \mathbb{R}^4 : \langle q, v\rangle = 0\}$. The tangent space of the product manifold $\mathcal{M} = (\mathbb{S}^3)^J$ at $x = (q^1, \dots, q^J)$ is the product $T_x\mathcal{M} = \prod_{j=1}^J T_{q^j}\mathbb{S}^3$, again a linear subspace of $\mathbb{R}^{4J}$. The exponential map on $\mathbb{S}^3$ admits the closed form

$$\mathrm{Exp}_q(\omega) = \cos(\|\omega\|)\, q + \frac{\sin(\|\omega\|)}{\|\omega\|}\,\omega, \qquad \omega \in T_q\mathbb{S}^3, \tag{47}$$

with $\mathrm{Exp}_q(0) = q$ by continuity. The exponential map on the product manifold $\mathcal{M}$ acts joint-wise through Eq. 47. The logarithmic map is its local inverse $\mathrm{Log}_q : \mathbb{S}^3 \to T_q\mathbb{S}^3$, defined whenever the two points are not antipodal. In particular, $\mathrm{Log}_{x_0}(x_1) \in T_{x_0}\mathcal{M}$ whenever $x_0, x_1 \in \mathcal{M}$ and no joint of x_1 is antipodal to the corresponding joint of x_0. The sign disambiguation of Appendix C.1 excludes this case.

C.4.2 Derivation of the Exactness Result

Setup. The correction network produces a raw output $\tilde{\epsilon}_\theta(t, x_0, x_1) \in \mathbb{R}^{4J}$, which is orthogonally projected onto the tangent space $T_{x_0}\mathcal{M}$ by

$$\epsilon_\theta(t, x_0, x_1) = P_{x_0}\,\tilde{\epsilon}_\theta(t, x_0, x_1), \qquad \big(P_{x_0}\,u\big)^j = u^j - \big\langle q_0^j,\, u^j\big\rangle\, q_0^j, \tag{48}$$

where $x_0 = (q_0^1, \dots, q_0^J)$ and u^j denotes the j-th joint component of $u \in \mathbb{R}^{4J}$. With this projection, the interpolation resides exactly on the motion manifold: for any endpoints $x_0, x_1 \in \mathcal{M}$, any raw correction $\tilde{\epsilon}_\theta(t, x_0, x_1) \in \mathbb{R}^{4J}$, and every $t \in [0,1]$, the interpolation of Eq. 11 satisfies $x_{t,\theta} \in \mathcal{M}$. We verify this in three steps.

Step 1: exact tangency of the correction. Fix a joint j. Since $x_0 \in \mathcal{M}$, we have $\|q_0^j\| = 1$, and hence for every raw output $\tilde{\epsilon}_\theta^j \in \mathbb{R}^4$,

$$\big\langle q_0^j,\ \tilde{\epsilon}_\theta^j - \big\langle q_0^j, \tilde{\epsilon}_\theta^j\big\rangle\, q_0^j\big\rangle = \big\langle q_0^j, \tilde{\epsilon}_\theta^j\big\rangle - \big\langle q_0^j, \tilde{\epsilon}_\theta^j\big\rangle\, \|q_0^j\|^2 = 0. \tag{49}$$

By the characterization of $T_q\mathbb{S}^3$ in Sec. C.4.1, this means $\epsilon_\theta^j \in T_{q_0^j}\mathbb{S}^3$, and therefore $\epsilon_\theta \in T_{x_0}\mathcal{M}$. This step uses no property of $\tilde{\epsilon}_\theta$: the raw network output is an arbitrary vector in the ambient space and is never required to lie on the manifold.

Step 2: exact tangency of the path argument. By definition of the logarithmic map, $\xi := \mathrm{Log}_{x_0}(x_1) \in T_{x_0}\mathcal{M}$, and $T_{x_0}\mathcal{M}$ is a linear subspace of $\mathbb{R}^{4J}$. Hence

$$\omega(t) := t\,\xi + t(1-t)\,\epsilon_\theta \in T_{x_0}\mathcal{M}, \qquad \forall\, t \in [0,1]. \tag{50}$$

Step 3: constraint preservation under the exponential map. Fix $t \in [0,1]$ and a joint j. Since $\langle q_0^j, \omega^j(t)\rangle = 0$ by Eq. 50, Eq. 47 yields

$$\|x_{t,\theta}^j\|^2 = \cos^2(\|\omega^j\|)\,\underbrace{\|q_0^j\|^2}_{=1} + 2\cos(\|\omega^j\|)\,\frac{\sin(\|\omega^j\|)}{\|\omega^j\|}\,\underbrace{\langle q_0^j, \omega^j\rangle}_{=0} + \frac{\sin^2(\|\omega^j\|)}{\|\omega^j\|^2}\,\|\omega^j\|^2 = 1, \tag{51}$$

where the degenerate case $\omega^j = 0$ follows by continuity ($\mathrm{Exp}_q(0) = q$ with $\|q\| = 1$). Therefore $x_{t,\theta}^j \in \mathbb{S}^3$ for every joint j, that is, $x_{t,\theta} \in \mathcal{M}$ for all $t \in [0,1]$.

Algorithm 2 Training and sampling of the sign motion generation network

Require: motion manifold $\mathcal{M}$, source and target pose distributions p_0, p_1, frozen correction network ϵ_θ, initialized generation network ϕ, text condition c

Phase 1: training ϕ along the naturalness-guided interpolation

while Training **do**
  Sample a training pair (x_1, e_1), sources $x_0 \sim p_0$, $e_0 \sim \mathcal{N}(0, \mathbf{I})$, and $t \sim \mathcal{U}[0,1]$
  $x_t = \mathrm{Exp}_{x_0}\big(t\,\mathrm{Log}_{x_0}(x_1) + t(1-t)\,\epsilon_\theta(t, x_0, x_1)\big)$ ▷ Eq. 11
  $e_t = (1-t)\,e_0 + t\,e_1$
  $(\tilde{x}_1, \hat{e}_1) = \phi(x_t, e_t, t, c)$, with $\hat{x}_1$ normalized per joint onto $\mathcal{M}$
  Update ϕ using gradient $\nabla_\phi\big(10\,\|x_1 - \hat{x}_1\|^2 + 0.1\,\|e_1 - \hat{e}_1\|^2\big)$ ▷ Eq. 13
end while

Phase 2: sampling with the path-consistent transport

$x_0 \sim p_0$, $e_0 \sim \mathcal{N}(0, \mathbf{I})$, $N \leftarrow 5$, $\Delta \leftarrow 1/N$ ▷ uniform grid $t_k = k\Delta$
for $k = 0, \ldots, N-1$ **do**
  $t \leftarrow k\Delta$; $(\tilde{x}_1, \hat{e}_1) = \phi(x_t, e_t, t, c)$
  $\hat{x}_1^j \leftarrow \tilde{x}_1^j / \|\tilde{x}_1^j\|$ for every joint j ▷ onto $\mathcal{M}$
  $x_{t+\Delta} = \mathrm{Exp}_{x_0}\big((t{+}\Delta)\,\mathrm{Log}_{x_0}(\hat{x}_1) + (t{+}\Delta)(1{-}t{-}\Delta)\,\epsilon_\theta(t{+}\Delta, x_0, \hat{x}_1)\big)$ ▷ Eq. 53
  $e_{t+\Delta} = e_t + \frac{\Delta}{1-t}\,(\hat{e}_1 - e_t)$
  renormalize $x_{t+\Delta}$ per joint onto $\mathcal{M}$ ▷ numerical safeguard
end for
return x_1, e_1

C.5 Training and Sampling the Generation Network

Training. We train ϕ with the target-prediction objective of Eq. 13 along the naturalness-guided interpolation. For a sampled source $x_0 \sim p_0$, target x_1, and time $t \sim \mathcal{U}[0,1]$, the training state is the path point of Eq. 11 with the frozen correction ϵ_θ. The network outputs a raw pose prediction $\tilde{x}_1 \in \mathbb{R}^{4J}$ and an expression prediction $\hat{e}_1$, and the pose prediction is normalized per joint onto $\mathcal{M}$ before the loss is computed:

$$\big(\tilde{x}_1, \hat{e}_1\big) = \phi\big(x_t, e_t, t, c\big), \qquad \hat{x}_1^j = \frac{\tilde{x}_1^j}{\big\|\tilde{x}_1^j\big\|}, \quad j = 1, \ldots, J, \tag{52}$$

where $\tilde{x}_1^j \in \mathbb{R}^4$ is the quaternion prediction of joint j. The loss of Eq. 13 is evaluated on the normalized prediction $\hat{x}_1 = (\hat{x}_1^1, \ldots, \hat{x}_1^J)$.

Inference. Since ϕ is trained to predict the target rather than the transport, we obtain the transport at sampling time by re-planning the naturalness-guided interpolation with the current prediction. At each solver step, we re-predict the target motion $\hat{x}_1 = \phi(x_t, e_t, t, c)$ from the current state and place the state at

$$x_{t+\Delta} = \mathrm{Exp}_{x_0}\Big(\underbrace{(t{+}\Delta)\,\mathrm{Log}_{x_0}\big(\hat{x}_1\big)}_{\text{Geometric}} + \underbrace{(t{+}\Delta)\big(1{-}t{-}\Delta\big)\,\epsilon_\theta\big(t{+}\Delta,\, x_0,\, \hat{x}_1\big)}_{\text{Naturalness}}\Big), \tag{53}$$

where x_0 is the retained source of the trajectory and the frozen correction ϵ_θ is projected onto $T_{x_0}\mathcal{M}$ as in Eq. 48. Eq. 53 evaluates the path of Eq. 11 for the pair $(x_0, \hat{x}_1)$.

Expression channel. The sign motion further contains facial expression parameters, which the manifold construction does not cover. We generate them alongside the joint rotations in Euclidean space: the generation network ϕ takes the paired expression state e_t as an additional input and predicts the target expression $\hat{e}_1$, trained under the same target-prediction objective as Eq. 13, with a loss weight of 10 for the pose term and 0.1 for the expression term. The expression source e_0 is Gaussian noise, and the expression path is the Euclidean interpolation $e_t = (1-t)\,e_0 + t\,e_1$. At sampling time, the expression state follows the analogous re-planned scheme, $e_{t+\Delta} = e_t + \frac{\Delta}{1-t}\,(\hat{e}_1 - e_t)$, which coincides with the linear path from e_0 to the re-predicted $\hat{e}_1$. A description of the complete training and sampling procedures of both channels is given in Algorithm 2.

D SUPPLEMENTARY EXPERIMENTS

D.1 EXPERIMENTAL DETAIL

Dataset Details. Phoenix-2014T is a German Sign Language (DGS) resource for weather forecasts, comprising 8,257 video sequences performed by 9 signers. CSL-Daily is a large-scale Chinese Sign Language (CSL) dataset focused on daily-life interactions, containing 20,654 clips recorded by 10 signers. How2Sign is an American Sign Language (ASL) dataset geared towards instructional videos, encompassing over 31,000 samples contributed by 11 signers. For sign motion extraction, we adopt SMPL-X (Pavlakos et al., 2019) parameters, which are sourced from NSA (Baltatzis et al., 2024) for How2Sign and from SOKE (Zuo et al., 2025) for Phoenix-2014T and CSL-Daily.

Table 6: Hyperparameters of the correction network ϵ_θ and the sign motion generation network ϕ.

	$\epsilon_\theta(t, x_0, x_1)$	$\phi(x_t, e_t, t, c)$
Architecture	3-layer MLP	8-layer transformer decoder
MLP hidden dimension	512	–
Latent dimension	–	512
Feed-forward dimension	–	2048
Attention heads	–	8
Time embedding dimension	–	2048
Dropout	–	0.1
Text conditioning	–	cross-attention to mBART features
Timestep injection	–	Stylization blocks
Batch size	8	64
Epochs	100	300
Optimizer	AdamW	AdamW
Learning rate	2e-4 (cosine)	2e-4 (cosine)
Weight decay	0.0	0.0
Precision	fp32	fp32
Motion channels: $J = 41$		
Expression dimension: $= 10$		
Sources: p_0 uniform on the sign-convention hemisphere; $e_0 \sim \mathcal{N}(0, \mathbf{I})$		
Loss weights: pose 10, expression 0.1		
Naturalness measure: $K = 500$, $\sigma = 0.50$, $\rho = 10^{-6}$		
Sampler: explicit Euler, 5 steps (NFE $= 5$)		

Implementation Details. Both the correction network ϵ_θ and the sign motion generation network ϕ are trained with the AdamW optimizer (Loshchilov & Hutter, 2017) under a cosine learning rate schedule starting at 2×10^{-4}. The correction network ϵ_θ is a 3-layer MLP with a hidden dimension of 512, trained with the kinetic-energy objective of Eq. 12 for 100 epochs with a batch size of 8. The generation network ϕ is a transformer decoder with 8 layers, trained with the target-prediction objective of Eq. 13 along the naturalness-guided interpolation for 300 epochs with a batch size of 64 per GPU. The text condition c is encoded by the pre-trained mBART-large-cc25 (Liu et al., 2020), which is fine-tuned jointly with ϕ, and the timestep t is injected into the decoder of ϕ through Stylization blocks (Zhang et al., 2024) after every self-attention, cross-attention, and feed-forward layer. All models are trained on 4 NVIDIA A40 GPUs.

The joint rotations are represented as J unit quaternions under the sign convention of Appendix C.1, and the facial expression parameters form an auxiliary Euclidean channel (Appendix C.5). The source distribution p_0 is uniform on the hemisphere fixed by the sign convention, and the expression source e_0 follows the standard Gaussian; training uses a loss weight of 10 for the pose term and 0.1 for the expression term. The naturalness measure uses $K = 500$ prototypes per joint with kernel bandwidth $\sigma = 0.50$ (Appendix C.2) and regularization constant $\rho = 10^{-6}$. Sampling integrates the re-anchored transport with the explicit Euler method in 5 steps (Appendix C.5). For SignNMFlow-M, the source-language token of mBART is set per sample according to the dataset of origin. Table 6 summarizes all hyperparameters.

Table 7: Efficiency comparison with state-of-the-art SLP methods. Latencies are measured using a single NVIDIA H100 GPU.

Methods	#Params (M)	NFE	Latency (s / motion)
SOKE	430.40	–	0.2687
SignFlow	261.04	25	0.0860
SignNMFlow	233.20	5	0.0103

Table 8: Ablation results of the number of function evaluations (NFE) in sampling. All results are evaluated on CSL-Daily.

NFE	DTW-JPE↓			DTW-PA-JPE↓		
	All	Body	Hand	All	Body	Hand
5	**16.86**	**5.73**	**7.43**	**8.92**	**5.10**	**1.30**
10	16.91	5.74	7.49	9.01	5.13	1.32
20	16.98	5.75	7.55	9.11	5.16	1.33

D.2 Additional Experimental Results

Number of Function Evaluations. In Table 8, we present the ablation results of the number of function evaluations (NFE) in sampling. Increasing NFE from 5 to 20 only varies the DTW-JPE All error from 16.86 to 16.98 and the DTW-PA-JPE All error from 8.92 to 9.11, with consistent behavior on body and hand joints. The sampling quality is thus robust to NFE. This accords with the re-planned scheme of Eq. 53, where each solver step re-predicts the target and places the state exactly at the corresponding path point, so the discretization of the solver has little influence on the generated motions. Since 5 steps already saturate the performance, we use NFE $= 5$ by default, which keeps sampling efficient.

Efficiency Comparison. In Table 7, we compare the efficiency of SignNMFlow with SOKE and SignFlow, where the latency is measured per sign motion on a single NVIDIA H100 GPU. SignNMFlow uses 233.20M parameters and attains a latency of 0.0103 seconds per motion, an $8.3\times$ speedup over SignFlow (261.04M parameters, 0.0860 seconds) and a $26.1\times$ speedup over SOKE (430.40M parameters, 0.2687 seconds). The advantage stems from two sources. First, the compact architecture keeps the model size small, whereas SOKE additionally incorporates a pre-trained language model for semantic enrichment. Second, the re-planned scheme of Eq. 53 requires only 5 function evaluations, whereas SignFlow takes 25, and Table 8 shows that these 5 evaluations already saturate the sampling quality. SignNMFlow therefore offers the best efficiency among the compared methods.