PHONONOISE: RESHAPING DIFFUSION GEOMETRY VIA PHONOLOGY-AWARE NON-ISOTROPIC DIFFUSION FOR SIGN LANGUAGE PRODUCTION

Anonymous authors
Paper under double-blind review

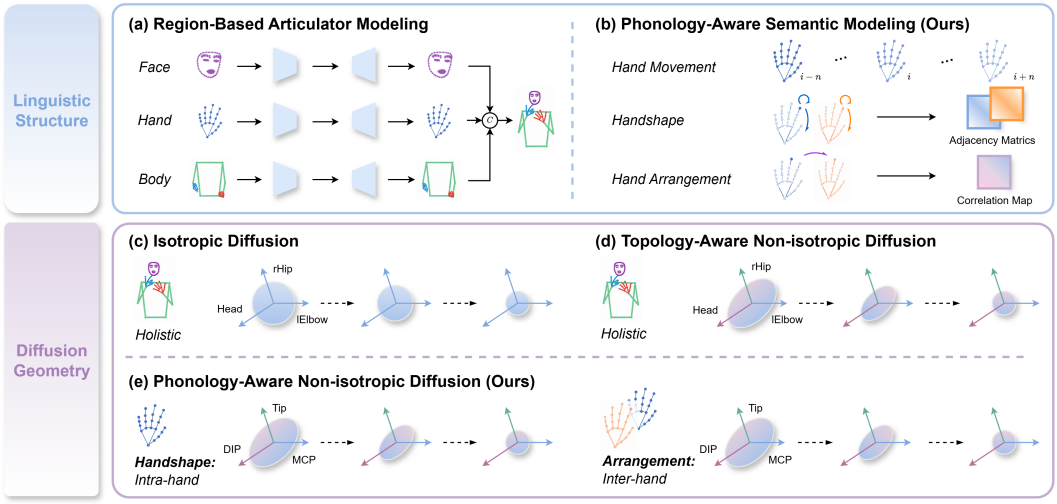

Figure 1: **Insight of our phonology-aware non-isotropic diffusion method *PhonoNoise* for sign language production. Linguistic structure (a–b):** Conventional methods organize sign motion according to anatomical regions, whereas we model linguistically grounded phonological factors. **Diffusion geometry (c–e):** Standard isotropic diffusion treats motion dimensions independently in the corruption process, while prior work on generic human motion prediction introduces non-isotropic diffusion based on skeletal topology. To the best of our knowledge, we are the first to ground non-isotropic hand corruption in sign phonology through hand relations.

ABSTRACT

Sign Language Production (SLP) aims to generate semantically faithful and articulatorily coherent sign motion from spoken-language text. Since the hands carry much of a sign's lexical content, errors in handshape, inter-hand configuration, and movement directly impair intelligibility, yet existing methods still often produce distorted fingers, stretched hand bones, and inaccurate trajectories. We attribute this partly to the forward process of diffusion-based methods, which corrupts all joints with isotropic noise and thus leaves hand-joint dependencies to be learned implicitly from data. Motivated by sign phonology, where handshape and inter-hand arrangement describe spatial hand relations and movement describes how the hands travel over time, we propose a phonology-aware diffusion framework with two components. PhonoNoise derives a phonology-informed covariance from intra-hand topology and bilateral joint correspondence for non-isotropic hand diffusion. PhonoContext extracts movement context from noisy wrist trajectories and injects it into the denoiser. Neither requires phonological annotations. Experiments on PHOENIX-2014T and CSL-Daily show state-of-the-art back-translation performance, with BLEU-4 scores of 17.11 and 6.21, respectively, while substantially reducing limb stretching. Code and more visualizations will be released on our project page.

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.

1 INTRODUCTION

Sign language is a visual language and a primary means of communication for many Deaf and Hard-of-Hearing (DHH) communities. Sign Language Production (SLP) aims to generate continuous sign sequences from spoken-language text, providing an automatic means of conveying spoken-language content through signing. In this sense, SLP is not merely motion generation but language generation through motion: successful signing requires both coherent articulated motion and accurate expression of the intended linguistic content. The hands are central to both, since much of a sign's lexical content is carried by handshape, the relation between the two hands, and hand movement. Two signs may differ only in handshape or movement, much like minimal pairs in spoken languages, so even small errors in these properties can change or obscure the intended meaning.

Recent SLP research has explored richer motion representations and more expressive generative models. Discrete latent spaces have been introduced to model continuous sign sequences (Xie et al., 2024; Yin et al., 2024), temporal representations have been improved (Qi et al., 2024), and diffusion-based approaches have incorporated anatomical graph or bone-level structure (Baltatzis et al., 2024; Tang et al., 2025a). Despite this progress, generated hands remain a weak point. As shown in Fig. 3, state-of-the-art methods still produce distorted finger configurations, incorrect inter-hand arrangements, and inaccurate movement trajectories. We identify two factors behind this limitation, one linguistic and one generative.

From the linguistic perspective, existing methods typically treat SLP as generic motion generation and organize structure anatomically, for example by separating body regions (Fig. 1(a–b)). Such modeling answers *which body region a joint belongs to*. Sign language studies, however, describe signs through phonological parameters such as handshape, movement, location, and orientation, together with non-manual cues from the face and body (Brentari, 1998; Adam et al., 2011; Cooper et al., 2012; Hong & Kosecka, 2026). Among them, handshape specifies coordinated configurations within each hand, hand arrangement specifies how the two hands are coordinated in two-handed signs, and movement describes how the hands travel over time. These parameters point to *which relations among articulators matter for conveying meaning*, a notion of structure that anatomical partitioning does not provide. They also differ in nature: handshape and arrangement are spatial relations among hand joints, whereas movement is a temporal property of hand trajectories, suggesting that they should enter a generative model in different ways.

From the generative perspective, structural priors in SLP have so far been introduced mainly through representations and network architectures, while the diffusion process itself remains unstructured. Standard Gaussian diffusion (Ho et al., 2020; Tevet et al., 2023) corrupts all motion dimensions with isotropic noise, treating hand joints as independent and leaving their dependencies to be learned implicitly from data. This is particularly problematic for the hands, which occupy a small region yet contain many tightly coupled joints: independent perturbations of neighboring finger joints quickly destroy bone lengths and finger configurations, and the denoiser must recover them without structural guidance. Recent work on human motion prediction (Curreli et al., 2025) has shown that joint correlations can instead be built into the diffusion process through non-isotropic noise derived from the skeletal graph. A whole-body skeletal graph, however, mixes body and hand connectivity and contains no relation between homologous joints of the two hands (Fig. 1(c–d)), leaving out precisely the hand relations emphasized by sign phonology. This raises the question: ***can sign phonology guide how the diffusion geometry is shaped?***

Motivated by these observations, we propose a phonology-aware diffusion framework that incorporates phonological factors into different parts of the generative process according to their spatial and temporal roles. For spatial phonology, PhonoNoise (Fig. 1(e)) instantiates handshape and inter-hand arrangement through intra-hand topology and bilateral joint correspondence, and converts them into a phonology-informed covariance for non-isotropic diffusion over the hand joints, while the body and face retain standard isotropic diffusion. For temporal phonology, PhonoContext extracts movement context from noisy wrist trajectories with a timestep-conditioned temporal network and injects it into the denoiser through a gated residual. Since movement unfolds over time rather than across joints, it is better supplied as context to the denoiser than encoded in a static covariance. In this design, phonologically motivated spatial relations shape *how hand motion is corrupted*, whereas movement informs *how corrupted motion is interpreted during denoising*. Neither component requires phonological annotations, so the framework can be trained on standard text-to-pose data.

Our contributions are summarized as follows:

- We present a phonology-aware diffusion formulation for SLP that uses phonologically motivated structure as an inductive bias without phonological annotations, building handshape and inter-hand arrangement into the diffusion process and supplying movement to the denoiser according to their spatial and temporal roles.
- We propose PhonoNoise, which derives a phonology-informed covariance from intra-hand topology and bilateral joint correspondence for non-isotropic hand diffusion while keeping the body and face isotropic, and PhonoContext, which injects movement-aware context from noisy wrist trajectories into the denoiser to capture temporal hand dynamics.
- Experiments on PHOENIX-2014T and CSL-Daily show that our method achieves state-of-the-art back-translation performance, and our ablation on both two datasets demonstrate that PhonoNoise substantially reduces limb stretching.

2 Related Work

Diffusion-based sign language production. SLP has evolved from autoregressive decoding (Saunders et al., 2020; Hwang et al., 2023; Saunders et al., 2021) to parallel non-autoregressive generation (Huang et al., 2021; Hwang et al., 2021; 2022; Taşyürek et al., 2026b; Kızıltepe et al., 2025; Taşyürek et al., 2026a), with recent work increasingly exploring iterative probabilistic generation such as diffusion models and conditional flow matching (Low et al., 2026). Within diffusion-based SLP, existing methods modify different components of the pipeline: G2P-DDM (Xie et al., 2024) and Sign-IDD (Tang et al., 2025a) redesign the motion representation space; MaDiS (Zuo et al., 2026) and SignPR (Liu et al., 2026) alter the generation strategy through masked or progressive block-wise diffusion; and GCDM (Tang et al., 2025b) and FGDM (Yu et al., 2026) enhance conditioning and denoising architectures. Other methods incorporate finer articulator-level conditions or adapt pretrained diffusion models for sign generation (Wang et al., 2026; Pelykh et al., 2026). Overall, diffusion-based SLP remains underexplored along two complementary dimensions: linguistically, existing methods largely rely on articulator-level or externally provided conditions rather than embedding phonological structure as an internal inductive bias; from the diffusion generative perspective, they primarily redesign representations, conditioning, generation strategies, or denoisers, while leaving the forward corruption process largely unchanged.

Phonology-aware sign language production. Sign languages are organized by sub-lexical articulatory properties such as handshape, location, movement, and inter-hand configuration, providing linguistic structure beyond gloss-level semantics (Adam et al., 2011). Existing SLP methods have begun to exploit such information, either through explicit phonological notation (Arkushin et al., 2023; Gil-Martín et al., 2023) or as additional generation conditions derived from annotated linguistic attributes (Hong & Kosecka, 2026). While these approaches demonstrate the value of phonological knowledge, they primarily treat it as externally provided symbolic input or conditioning. In contrast, we treat phonological principles as internal structural inductive biases that shape both the diffusion geometry and the denoising representation.

Structured Noise in Diffusion Models. Standard diffusion models typically adopt i.i.d. isotropic Gaussian noise (Ho et al., 2020), while recent studies introduce structured corruption across spatial (Yu et al., 2023), temporal (Ge et al., 2023; Chang et al., 2024), and spectral domains (Rissanen et al., 2022; Hoogeboom & Salimans, 2022; Huang et al., 2024; Scimeca et al., 2025), or learn input-adaptive noise schedules (Sahoo et al., 2024). More recently, domain-specific priors have been incorporated into the forward process: SkeletonDiffusion (Curreli et al., 2025) derives non-isotropic correlations from skeletal kinematics, while edge-aware diffusion exploits image structures (Vandersanden et al., 2024). However, these priors remain primarily geometric or perceptual. Although phonological attributes have been explored as conditions for SLP (Hong & Kosecka, 2026), they have not been used to structure diffusion noise. To the best of our knowledge, we are the first to derive non-isotropic correlations from sign phonology, directly encoding handshape and inter-hand arrangement into the forward corruption process.

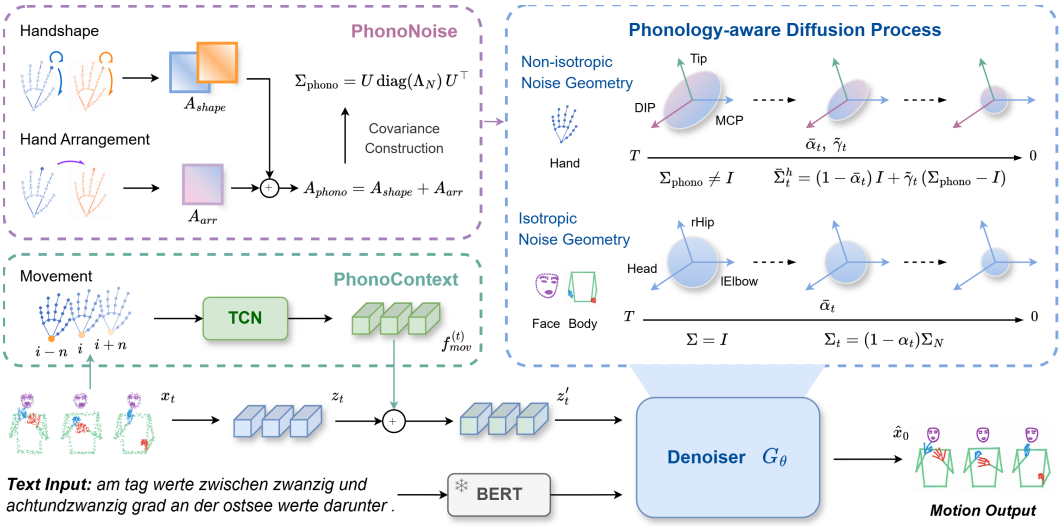

Figure 2: Overview of our phonology-aware diffusion framework for sign language production. **PhonoNoise** converts intra-hand handshape and inter-hand arrangement relations into a phonology-informed covariance, yielding non-isotropic diffusion for the hands while retaining isotropic diffusion for the body and face. **PhonoContext** extracts movement-aware context from noisy wrist trajectories and injects it into the pose representation as a residual. The resulting representation, together with the text condition, is processed by the denoiser to reconstruct the clean sign motion.

3 Method

3.1 Problem Formulation

Given a spoken-language sentence $\mathbf{X} = \{x_1, x_2, \ldots, x_W\}$ with W tokens, our goal is to generate a 3D sign pose sequence $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N\}$ with N frames, where $\mathbf{p}_n \in \mathbb{R}^{K\times 3}$ denotes the 3D coordinates of K body, hand, and facial joints. We denote the clean pose sequence by $\mathbf{x}_0$, with body, hand, and face components $\mathbf{x}_0^b$, $\mathbf{x}_0^h$, and $\mathbf{x}_0^f$, respectively.

3.2 Preliminaries: Standard Isotropic Motion Diffusion

Given a clean motion sequence $\mathbf{x}_0$, standard diffusion progressively corrupts it with isotropic Gaussian noise (Ho et al., 2020). Let β_t denote the variance schedule, $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. At diffusion timestep $t \in \{1, \ldots, T\}$, the marginal forward distribution is

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0,\ (1-\bar{\alpha}_t)\mathbf{I}\right). \tag{1}$$

Equivalently, the noisy motion can be sampled as

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{2}$$

The identity covariance $\mathbf{I}$ perturbs all motion dimensions independently with equal variance and thus encodes no dependencies among joints.

3.3 Phonology-Aware Diffusion Framework

We build our framework upon the human motion diffusion model (MDM) (Tevet et al., 2023). Given an input sentence $\mathbf{X}$, a frozen BERT encoder produces the text condition $\mathbf{c}$. Following the $\mathbf{x}_0$-prediction parameterization of MDM, the denoiser G_θ predicts the clean motion from the noisy motion $\mathbf{x}_t$, diffusion timestep t, and text condition $\mathbf{c}$:

$$\hat{\mathbf{x}}_0 = G_\theta(\mathbf{x}_t, t, \mathbf{c}). \tag{3}$$

As illustrated in Fig. 2, we incorporate sign phonology according to its spatial and temporal roles. For spatial phonology, **PhonoNoise** (Section 3.3.1) models intra-hand handshape and inter-hand arrangement relations to construct a phonology-informed covariance, which introduces non-isotropic corruption to the hand component while retaining isotropic diffusion for the body and face. The resulting forward process induces the corresponding non-isotropic reverse transitions. For temporal phonology, **PhonoContext** (Section 3.3.2) extracts movement-aware context from the noisy wrist trajectories and injects it into the denoiser as a residual representation. PhonoNoise thus acts on the diffusion process, whereas PhonoContext acts on the denoiser input.

3.3.1 PHONONOISE: SPATIAL PHONOLOGY FOR NON-ISOTROPIC DIFFUSION

Isotropic noise perturbs each hand joint independently, although handshape arises from coordinated configurations within each hand and hand arrangement from relations between the two hands. PhonoNoise builds these relations into the diffusion process: we construct relation matrices for handshape and hand arrangement, convert them into a covariance, and use it to corrupt the hand joints, so that the injected noise, and in turn the reverse transitions, follow spatial phonological structure.

A. Intra-Hand Handshape Relations. Handshape is characterized by coordinated configurations of joints within each hand. To encode this intra-hand structure, we construct fixed topology-based adjacency matrices for the two hands. Let $\mathbf{A}_R, \mathbf{A}_L \in \{0,1\}^{J_h \times J_h}$ denote the binary adjacency matrices of the right and left hands, respectively, where an entry is nonzero if two joints are connected according to the hand kinematic topology. We then define the handshape adjacency matrix as

$$\mathbf{A}_{\text{shape}} = \begin{bmatrix} \mathbf{A}_R & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_L \end{bmatrix}, \tag{4}$$

where J_h is the number of joints per hand. The block-diagonal structure captures intra-hand dependencies while excluding cross-hand interactions, which are modeled via inter-hand arrangement.

B. Inter-Hand Arrangement Relations. Inter-hand arrangement describes the relative organization of the two hands and therefore requires modeling dependencies across hands. We establish correspondences between homologous joints of the right and left hands and define a binary correspondence matrix $\mathbf{B}_{RL} \in \{0,1\}^{J_h \times J_h}$, where $\mathbf{B}_{RL}(i,j) = 1$ if the i-th right-hand joint corresponds to the j-th left-hand joint. The resulting inter-hand arrangement relation matrix is

$$\mathbf{A}_{\text{arr}} = \begin{bmatrix} \mathbf{0} & \mathbf{B}_{RL} \\ \mathbf{B}_{RL}^\top & \mathbf{0} \end{bmatrix}. \tag{5}$$

$\mathbf{A}_{\text{arr}}$ thus links each joint to its counterpart on the other hand, which $\mathbf{A}_{\text{shape}}$ does not capture.

C. Phonology-Induced Noise Geometry. Based on the above designs, we then convert the two relation matrices into the noise geometry used for hand diffusion.

phonology-informed covariance. We combine the intra- and inter-hand structures as $\mathbf{A}_{\text{phono}} = \mathbf{A}_{\text{shape}} + \mathbf{A}_{\text{arr}}$. Since the resulting relation matrix is not guaranteed to define a valid covariance, we shift its spectrum and normalize its spectral scale:

$$\mathbf{\Sigma}_{\text{phono}} = \frac{\mathbf{A}_{\text{phono}} + c\mathbf{I}}{\lambda_{\max}(\mathbf{A}_{\text{phono}} + c\mathbf{I})}, \qquad c = \rho(\mathbf{A}_{\text{phono}}) + \varepsilon, \tag{6}$$

where $\rho(\cdot)$ denotes the spectral radius and $\varepsilon > 0$ is a small constant. By construction, all eigenvalues of $\mathbf{\Sigma}_{\text{phono}}$ lie in $(0, 1]$, so it is a valid positive-definite covariance.

Non-isotropic covariance schedule. Rather than applying the phonological covariance uniformly throughout diffusion, we interpolate between isotropic and phonology-induced noise using a timestep-dependent coefficient $\gamma_t \in [0, 1]$. The single-step covariance of the hand component is

$$\mathbf{\Sigma}_t^h = \beta_t \left[(1-\gamma_t)\mathbf{I} + \gamma_t \mathbf{\Sigma}_{\text{phono}}\right], \tag{7}$$

where, following (Curreli et al., 2025), we set $\gamma_t = \beta_t$ under a cosine schedule (Nichol & Dhariwal, 2021), so that the phonological structure mainly takes effect at high noise levels. The corresponding cumulative hand covariance is $\bar{\mathbf{\Sigma}}_t^h = (1-\bar{\alpha}_t)\mathbf{I} + \tilde{\gamma}_t(\mathbf{\Sigma}_{\text{phono}} - \mathbf{I})$, where $\tilde{\gamma}_t = \bar{\alpha}_t \sum_{\tau=1}^{t} \beta_\tau \gamma_\tau / \bar{\alpha}_\tau$.

Region-wise forward and reverse diffusion. PhonoNoise modifies only the hand corruption process, while the body and face retain the standard isotropic diffusion. We therefore define the cumulative covariance over the joint regions as

$$\bar{\mathbf{\Sigma}}_t^{\text{pose}} = \text{BlkDiag}\left((1-\bar{\alpha}_t)\mathbf{I}_b, \bar{\mathbf{\Sigma}}_t^h, (1-\bar{\alpha}_t)\mathbf{I}_f\right), \tag{8}$$

where $\mathbf{I}_b$ and $\mathbf{I}_f$ denote identity matrices for the body and face joints, respectively. For clarity, Eq. 8 describes the covariance over joint dimensions; the same structure is applied independently across spatial coordinates and motion frames. The resulting forward process is

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0,\, \bar{\mathbf{\Sigma}}_t^{\text{pose}}\right). \tag{9}$$

Since the process remains linear Gaussian, the posterior $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)$ is available in closed form (Curreli et al., 2025). During reverse sampling, we use the non-isotropic posterior for the hand component and the standard isotropic posterior for body and face.

3.3.2 PHONOCONTEXT: MOVEMENT-AWARE TEMPORAL CONTEXT

Unlike handshape and hand arrangement, movement is defined by how the hands travel over time. We furtherly design PhonoContext that extracts movement-aware context directly from the noisy motion $\mathbf{x}_t$ and injects it into the denoiser representation.

For each hand $s \in \{L, R\}$ and frame n, we represent movement using both the wrist position and its displacement relative to an upper-body reference:

$$\mathbf{u}_{s,n}^{(t)} = \left[\mathbf{w}_{s,n}^{(t)}; \mathbf{w}_{s,n}^{(t)} - \mathbf{o}_n^{(t)}\right], \tag{10}$$

where $\mathbf{w}_{s,n}^{(t)}$ is the wrist position from $\mathbf{x}_t$, and $\mathbf{o}_n^{(t)}$ is the shoulder-center reference. The first term preserves the wrist trajectory in pose coordinates, while the second captures body-relative motion.

The bilateral movement sequences $\{\mathbf{u}_{s,n}^{(t)}\}$ are processed by a shared timestep-conditioned temporal convolutional network (TCN) to capture their temporal dynamics. The resulting left- and right-hand features are then fused into a movement context $\mathbf{f}_{\text{mov}}^{(t)}$. Let $\mathbf{z}^{(t)}$ denote the frame-wise pose embedding of $\mathbf{x}_t$. We inject the movement context as a gated residual:

$$\tilde{\mathbf{z}}^{(t)} = \mathbf{z}^{(t)} + \tanh(g)\, \mathbf{f}_{\text{mov}}^{(t)}, \tag{11}$$

where g is a learnable scalar gate initialized at zero. The augmented representation $\tilde{\mathbf{z}}^{(t)}$ is then processed by the motion denoiser together with the text condition. PhonoContext leaves the diffusion covariance unchanged and requires no movement annotations.

3.3.3 TRAINING OBJECTIVE

Following the $\mathbf{x}_0$-prediction formulation, we align the reconstruction objective with the region-wise diffusion geometry above. For the isotropic body and face regions, we use mean-squared reconstruction losses. For the hands, we follow the covariance-aware formulation of Curreli et al. (2025) and measure residuals in the Mahalanobis-whitened space induced by PhonoNoise.

Let $\mathbf{\Sigma}_{\text{phono}} = \mathbf{U}\,\text{diag}(\boldsymbol{\lambda})\mathbf{U}^\top$ be its eigendecomposition. Since $\bar{\mathbf{\Sigma}}_t^h$ is a linear combination of $\mathbf{I}$ and $\mathbf{\Sigma}_{\text{phono}}$, it shares the eigenvectors $\mathbf{U}$, i.e., $\bar{\mathbf{\Sigma}}_t^h = \mathbf{U}\,\text{diag}(\bar{\mathbf{\Lambda}}_t)\mathbf{U}^\top$ with $\bar{\mathbf{\Lambda}}_t = (1-\bar{\alpha}_t)\mathbf{1} + \tilde{\gamma}_t(\boldsymbol{\lambda} - \mathbf{1})$. Given the hand reconstruction residual $\mathbf{e}_t^h = \hat{\mathbf{x}}_0^h - \mathbf{x}_0^h$, we define hand reconstruction objective as

$$\mathcal{L}_{\text{Maha}}^h = \left\| \text{diag}\left(\bar{\mathbf{\Lambda}}_t^{-1/2}\right) \mathbf{U}^\top \mathbf{e}_t^h \right\|_1, \tag{12}$$

where the whitening reweights reconstruction errors according to the same non-isotropic geometry used in the forward diffusion process.

The overall reconstruction objective is

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{MSE}}^b + \mathcal{L}_{\text{Maha}}^h + \mathcal{L}_{\text{MSE}}^f, \tag{13}$$

where the superscripts b, h, and f denote the body, hand, and face regions, respectively, and all losses are computed over valid motion frames.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets. We conduct experiments on two widely used sign language datasets that differ in language and domain to assess the generalization of our method. PHOENIX-2014T (Camgoz et al., 2018) contains 8,257 German Sign Language videos performed by 9 signers, collected from weather forecast broadcasts with a German vocabulary of 2,887 words. CSL-Daily (Zhou et al., 2021) is a larger-scale Chinese Sign Language dataset of 20,654 videos covering daily communication scenarios, recorded by 10 signers with a Chinese vocabulary of 2,343 words.

Evaluation Metrics. We evaluate both semantic consistency and pose quality. For semantic consistency, we follow the protocol (Walsh et al., 2025) for German dataset, and follow many SLP works (Saunders et al., 2020) to train a Joint-SLT model (Camgoz et al., 2020) for Chinese dataset ourselves to back-translate generated poses into text and report BLEU, ROUGE, and chrF. For pose quality, we use DTW-MJE, which computes the mean joint error after dynamic time warping alignment. In the ablation study, we additionally report Stretch-Mean and Stretch-RMSE (Curreli et al., 2025), which measure hand bone-length deviation relative to the mean bone lengths of each GT sequence. We abbreviate DTW-MJE, BLEU, ROUGE, Stretch-Mean, and Stretch-RMSE as DTW, B, R, Str-Mean, and Str-RMSE, respectively.

Implementation Details. For Phoenix-2014T, we use the 3D 178-keypoint pose data from the SLRTP CVPR 2025 Challenge (Walsh et al., 2025). For CSL-Daily, we use the MSRSLP lift3d data (Zhu, 2026), obtained by lifting MediaPipe Holistic (Lugaresi et al., 2019) sequences to 3D with the model of Ivashechkin et al. (2023) and represented with 178 keypoints. We use a batch size of 32 for both datasets and conduct all experiments on a single NVIDIA V100 GPU.

4.1.1 COMPARISON WITH STATE-OF-THE-ART

Quantitative Results on Phoenix-2014T Dataset. As shown in Table 1, PhonoNoise achieves the strongest semantic performance on Phoenix-2014T. Compared with A^2V-SLP (Taşyürek et al., 2026a), the previous best-performing method on language-level metrics, PhonoNoise improves all BLEU scores, with B4 increasing from 13.31 to 17.11 (+3.80). It also improves R from 36.64 to 42.58 (+5.94) and chrF from 35.41 to 39.20 (+3.79). These consistent gains across multiple language-level metrics indicate that incorporating phonological structure into the diffusion process improves the semantic fidelity of generated signing motions. PhonoNoise obtains a DTW of 0.042, which remains competitive with prior methods although it is not the lowest, suggesting that geometric similarity alone may not fully reflect semantic preservation in sign language production.

Table 1: Comparison results on Phoenix-2014T. "–" indicates results not reported in original paper.

Model	Venue	DTW↓	B1↑	B2↑	B3↑	B4↑	R↑	chrF↑
GT		0.000	34.43	22.08	16.13	12.81	35.22	34.62
PT (Saunders et al., 2020)	ECCV20	0.191	11.45	7.08	5.08	4.04	14.87	19.09
G2P-DDM (Xie et al., 2024)	AAAI24	0.116	16.11	11.37	9.22	7.50	–	–
SLRTP25-Winner (Zuo et al., 2024)	ECCV24	0.045	34.85	21.96	15.65	12.06	36.59	36.83
Sign-IDD (Tang et al., 2025a)	AAAI25	–	24.80	–	–	9.08	26.58	–
ILRSLP (Kızıltepe et al., 2025)	ICCVW25	0.037	32.72	20.69	14.78	11.37	34.10	34.18
DARSLP (Taşyürek et al., 2026b)	WACV26	<u>0.039</u>	32.33	19.84	14.07	10.90	31.80	32.00
A^2V-SLP (Taşyürek et al., 2026a)	ArXiv26	**0.037**	<u>36.25</u>	<u>23.23</u>	<u>16.88</u>	<u>13.31</u>	<u>36.64</u>	<u>35.41</u>
FGDM (Yu et al., 2026)	CVPR26	–	26.51	17.68	12.71	9.67	28.45	–
SignPR (Liu et al., 2026)	CVPR26	0.111	31.91	–	–	9.41	32.86	–
PhonoNoise		0.042	**40.75**	**28.22**	**21.27**	**17.11**	**42.58**	**39.20**

Quantitative Results on CSL-Daily Dataset. As shown in Table 2, PhonoNoise achieves the best performance across all reported metrics on CSL-Daily. Compared with DARSLP (Taşyürek et al., 2026b), the strongest prior method on semantic metrics, PhonoNoise consistently improves all BLEU scores, with B4 increasing from 3.53 to 6.21 (+2.68). It further improves R from 20.55 to 24.81 (+4.26) and chrF from 4.89 to 6.78 (+1.89). In addition to the semantic improvements,

PhonoNoise reduces DTW from the previous best result of 0.108 achieved by SignPR to 0.077 (-0.031). The consistent improvements in both geometric and text-based metrics demonstrate the effectiveness of phonology-structured diffusion across different sign language datasets.

Table 2: Comparison results on CSL-Daily. "–" indicates results not reported in original paper.

Model	Venue	DTW↓	B1↑	B2↑	B3↑	B4↑	R↑	chrF↑
GT		0.000	23.06	11.90	7.33	5.04	22.39	5.88
PT (Saunders et al., 2020)	ECCV20	0.084	11.69	2.16	0.67	0.21	11.50	2.08
ILRSLP (Kızıltepe et al., 2025)	ICCVW25	0.179	19.80	9.24	5.13	3.12	19.53	4.49
DARSLP (Taşyürek et al., 2026b)	WACV26	0.175	20.47	9.55	5.45	3.53	20.55	4.89
A^2V-SLP (Taşyürek et al., 2026a)	ArXiv26	0.169	18.75	8.37	4.79	3.15	18.18	4.29
SignPR (Liu et al., 2026)	CVPR26	0.108	14.08	–	–	3.01	14.43	–
PhonoNoise		**0.077**	**24.71**	**13.81**	**8.86**	**6.21**	**24.81**	**6.78**

Qualitative Results on CSL-Daily Dataset. Figure 3 provides a comparison of phonological aspects of the generated signs. For movement, PhonoNoise better reproduces the global arm trajectory of the GT, while DARSLP deviates from the reference motion. The enlarged hand regions further reveal that this improvement is not limited to coarse dynamics. PhonoNoise retains more faithful handshapes and better preserves the spatial configuration between the two hands, whereas DARSLP exhibits distorted finger articulation and less accurate hand arrangement. The results visually support our motivation that explicitly structuring the diffusion process with phonological relations can benefit both coarse movement patterns and fine-grained manual details.

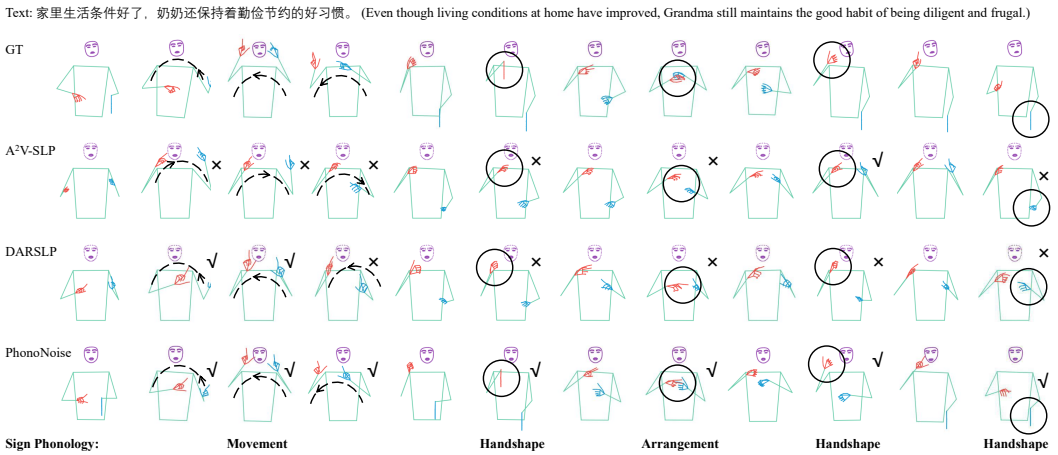

Figure 3: Qualitative comparison with state-of-the-art methods on CSL-Daily. Arrows indicate hand movement directions and circles highlight handshape and inter-hand arrangement; ✓ and × mark whether each phonological cue matches the ground truth.

4.1.2 Ablation Studies

Quantitative Results on Phoenix-2014T and CSL-Daily Datasets. Table 3 shows that PhonoNoise mainly benefits hand structure: it reduces Str-Mean from 9.07 to 5.38 on Phoenix-2014T and from 17.64 to 1.90 on CSL-Daily, and also raises all back-translation scores. PhonoContext brings smaller semantic gains (e.g., B4 from 15.77 to 16.16 on Phoenix-2014T and from 4.39 to 5.09 on CSL-Daily) but has little effect on the stretch metrics. With both components, the Full model is best on every metric on Phoenix-2014T, with a Str-RMSE (18.81) close to that of GT (17.34), and obtains the best semantic scores on CSL-Daily while staying close to PhonoNoise on hand structure. The structural gains are larger on CSL-Daily, likely because its GT hands keep constant bone lengths within a sequence, whereas the estimated hand keypoints of Phoenix-2014T fluctuate from frame to frame. Some models score above GT in back-translation; since the back-translation model is trained

on noisy GT poses, it decodes the smoother generated poses more easily, so these scores indicate decodability rather than fidelity beyond real signing.

Table 3: Ablation study on Phoenix-2014T and CSL-Daily. PN, PC, and Full denote PhonoNoise, PhonoContext, and the full model.

	Phoenix-2014T								CSL-Daily							
Design	Str-Mean↓	Str-RMSE↓	B1↑	B2↑	B3↑	B4↑	R↑	chrF↑	Str-Mean↓	Str-RMSE↓	B1↑	B2↑	B3↑	B4↑	R↑	chrF↑
GT	0.00	17.34	34.43	22.08	16.13	12.81	35.22	34.62	0.00	0.00	23.06	11.90	7.33	5.04	22.39	5.88
Baseline	9.07	28.68	40.66	27.42	20.10	15.77	42.03	38.32	17.64	23.21	21.55	10.89	6.49	4.39	21.26	5.39
+PN	5.38	18.98	40.74	28.01	21.03	16.85	42.20	39.07	**1.90**	**5.07**	24.46	13.62	8.65	6.04	24.53	6.69
+PC	9.74	26.49	40.70	27.59	20.38	16.16	42.06	38.97	16.28	21.74	22.73	11.88	7.35	5.09	22.71	5.93
+Full	**5.16**	**18.81**	**40.75**	**28.22**	**21.27**	**17.11**	**42.58**	**39.20**	**1.90**	5.28	**24.71**	**13.81**	**8.86**	**6.21**	**24.81**	**6.78**

Qualitative Results on CSL-Daily Dataset. Fig. 4 compares generations with and without each component. In the right example, PhonoNoise produces handshapes and inter-hand arrangements close to the ground truth, whereas without it the fingers are distorted and the two hands are misplaced relative to each other. In the left example, PhonoContext keeps the hand trajectories and movement directions consistent with the ground truth; without it, several movements are missing or go in the opposite direction. PhonoNoise thus mainly corrects the spatial configuration of the hands, while PhonoContext corrects how they move over time.

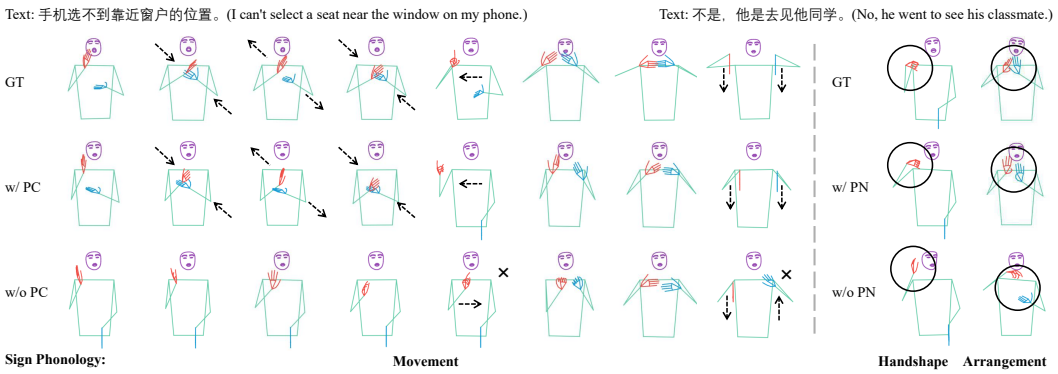

Figure 4: Qualitative ablation on CSL-Daily. PhonoNoise (PN) improves handshape and inter-hand arrangement, while PhonoContext (PC) better preserves hand movement trajectories and directions.

5 Conclusion

We presented a phonology-aware diffusion framework for sign language production, with structural priors motivated by the spatial and temporal roles of sign phonology. PhonoNoise instantiates handshape and arrangement relations through hand topology and bilateral joint correspondence to derive a phonology-informed covariance for non-isotropic hand diffusion, while PhonoContext supplies movement context from wrist trajectories to the denoiser. Experiments on PHOENIX-2014T and CSL-Daily show that the two components jointly improve the structural realism and semantic consistency of generated signs. For sign language production, our work suggests that phonologically motivated hand relations can serve as a structural prior on how sign motion is corrupted and recovered, without requiring phonological labels. More broadly, our results suggest that domain structure can be encoded in the noise geometry of diffusion models rather than only in their conditioning or architecture. A limitation is that our relations are instantiated through hand topology and bilateral correspondence rather than richer phonological categories, so whether the gains stem from phonology itself or from structured hand modeling in general remains to be examined. In future work, we will explore further phonological parameters such as location and non-manual markers, and develop evaluation protocols that better reflect fidelity to real signing.

AI USE STATEMENT

We used generative AI tools to aid language polishing. All AI-assisted outputs were reviewed by the authors, who take full responsibility for the final content of this work.

REFERENCES

Robert Adam, Eleni Orfanidou, James M McQueen, and Gary Morgan. *Sign language comprehension: Insights from misperceptions of different phonological parameters*. Walter de Gruyter, 2011.

Rotem Shalev Arkushin, Amit Moryossef, and Ohad Fried. Ham2pose: Animating sign language notation into pose sequences. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21046–21056. IEEE, 2023.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1985–1995. IEEE, 2024.

Diane Brentari. *A prosodic model of sign language phonology*. Mit Press, 1998.

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

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10023–10033, 2020.

Pascal Chang, Jingwei Tang, Markus Gross, and Vinicius Da Costa De Azevedo. How i warped your noise: a temporally-correlated noise prior for diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 30015–30046, 2024.

Helen Cooper, Eng-Jon Ong, Nicolas Pugeault, and Richard Bowden. Sign language recognition using sub-units. *The Journal of Machine Learning Research*, 13(1):2205–2231, 2012.

Cecilia Curreli, Dominik Muhle, Abhishek Saroha, Zhenzhang Ye, Riccardo Marin, and Daniel Cremers. Nonisotropic gaussian diffusion for realistic 3d human motion prediction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1871–1882. IEEE, 2025.

Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22873–22884. IEEE, 2023.

Manuel Gil-Martín, María Villa-Monedero, Andrzej Pomirski, Daniel Sáez-Trigueros, and Rubén San-Segundo. Sign language motion generation from sign characteristics. *Sensors*, 23(23):9365, 2023.

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

Rui Hong and Jana Kosecka. Toward phonology-guided sign language motion generation: A diffusion baseline and conditioning analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 10641–10648, June 2026.

Emiel Hoogeboom and Tim Salimans. Blurring diffusion models. *arXiv preprint arXiv:2209.05557*, 2022.

Wencan Huang, Wenwen Pan, Zhou Zhao, and Qi Tian. Towards fast and high-quality sign language production. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3172–3181, 2021.

Xingchang Huang, Corentin Salaun, Cristina Vasconcelos, Christian Theobalt, Cengiz Oztireli, and Gurprit Singh. Blue noise for diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

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

Eui Jun Hwang, Jung Ho Kim, Suk Min Cho, and Jong C Park. Non-autoregressive sign language production via knowledge distillation. *arXiv preprint arXiv:2208.06183*, 2022.

Eui Jun Hwang, Huije Lee, and Jong C Park. Autoregressive sign language production: A gloss-free approach with discrete representations. *arXiv preprint arXiv:2309.12179*, 2023.

Maksym Ivashechkin, Oscar Mendez, and Richard Bowden. Improving 3d pose estimation for sign language. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pp. 1–5. IEEE, 2023.

Tuğçe Kızıltepe, Sümeyye Meryem Taşyürek, and Hacer Yalim Keles. Iterative latent refinement for robust non-autoregressive sign language production. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4883–4893, 2025.

Xiao Liu, Shiwei Gan, Yafeng Yin, Bowen Guo, Zhiwei Jiang, Shunmei Meng, Lei Xie, and Sanglu Lu. Signpr: A progressive vector-quantized diffusion framework for sign language production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2198–2208, 2026.

Jianhe Low, Alexandre Symeonidis-Herzig, Maksym Ivashechkin, Özge Mercanoğlu Sincan, and Richard Bowden. Signspark: Efficient multilingual sign language production via sparse keyframe learning. In *European Conference on Computer Vision (ECCV)*, 2026.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PmLR, 2021.

Anton Pelykh, Edward Fish, Ozge Mercanoglu Sincan, and Richard Bowden. Signrefine: Adapting foundational video models for sign language generation. *arXiv preprint arXiv:2609.08496*, 2026.

Fan Qi, Yu Duan, Huaiwen Zhang, and Changsheng Xu. Signgen: End-to-end sign language video generation with latent diffusion. In *European Conference on Computer Vision*, pp. 252–270. Springer, 2024.

Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation. *arXiv preprint arXiv:2206.13397*, 2022.

Subham S Sahoo, Aaron Gokaslan, Chris De, and Volodymyr Kuleshov. Diffusion models with learned adaptive noise. *Advances in Neural Information Processing Systems*, 37:105730–105779, 2024.

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

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

Luca Scimeca, Thomas Jiralerspong, Berton Earnshaw, Jason Hartford, and Yoshua Bengio. Learning what matters: Steering diffusion via spectrally anisotropic forward noise. *arXiv preprint arXiv:2510.09660*, 2025.

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

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

Sümeyye Meryem Taşyürek, Enis Mücahid İskender, and Hacer Yalim Keles. A^2v-slp: Alignment-aware variational modeling for disentangled sign language production. *arXiv preprint arXiv:2602.11861*, 2026a.

Sümeyye Meryem Taşyürek, Tuğçe Kızıltepe, and Hacer Yalim Keles. Disentangle and regularize: Sign language production with articulator-based disentanglement and channel-aware regularization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 8458–8467, 2026b.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SJ1kSyO2jwu.

Jente Vandersanden, Sascha Holl, Xingchang Huang, and Gurprit Singh. Edge-preserving noise for diffusion models. In *Computer Graphics Forum*, pp. e70383. Wiley Online Library, 2024.

Harry Walsh, Ed Fish, Ozge Mercanoglu Sincan, Mohamed Ilyes Lakhal, Richard Bowden, Neil Fox, Bencie Woll, Kepeng Wu, Zecheng Li, Weichao Zhao, et al. Slrtp2025 sign language production challenge: Methodology, results and future work. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4148–4158, 2025.

Cong Wang, Zexuan Deng, Zhiwei Jiang, Yafeng Yin, Fei Shen, Zifeng Cheng, Shiping Ge, Shiwei Gan, and Qing Gu. Advanced sign language video generation with compressed and quantized multi-condition tokenization. *Advances in Neural Information Processing Systems*, 38:79519–79545, 2026.

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

Aoxiong Yin, Haoyuan Li, Kai Shen, Siliang Tang, and Yueting Zhuang. T2s-gpt: Dynamic vector quantization for autoregressive sign language production from text. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3345–3356, 2024.

Xi Yu, Xiang Gu, Haozhi Liu, and Jian Sun. Constructing non-isotropic gaussian diffusion model using isotropic gaussian diffusion model for image editing. *Advances in Neural Information Processing Systems*, 36:76839–76851, 2023.

Yiheng Yu, Sheng Liu, Yuan Feng, Zhelun Jin, Yining Jiang, and Min Xu. Focal-general diffusion model with semantic consistent guidance for sign language production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35915–35925, 2026.

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

Chengyao Zhu. Multi-stream residual quantization for sign language production, 2026. URL https://github.com/Larmiii/MSRSLP. GitHub repository.

Ronglai Zuo, Fangyun Wei, Zenggui Chen, Brian Mak, Jiaolong Yang, and Xin Tong. A simple baseline for spoken language to sign language translation with 3d avatars. In *European Conference on Computer Vision*, pp. 36–54. Springer, 2024.

Ronglai Zuo, Rolandos Alexandros Potamias, Qi Sun, Evangelos Ververas, Jiankang Deng, and Stefanos Zafeiriou. Madis: Taming masked diffusion language models for sign language generation. *arXiv preprint arXiv:2601.19577*, 2026.

A APPENDIX

A.1 ANALYSIS OF NON-ISOTROPIC SCHEDULING.

We further investigate how the strength of non-isotropic corruption should vary across diffusion timesteps under the full model. Following SkeletonDiffusion (Curreli et al., 2025), we use a timestep-dependent coefficient γ_t to control the contribution of the non-isotropic covariance. For a fair comparison, all variants use the same cosine β_t diffusion schedule and the same reconstruction objective: the hand joints are optimized with the WL1 Mahalanobis loss, while the body and face retain the standard MSE loss. Thus, the only difference among these experiments is the scheduling of γ_t.

We first consider a *constant* schedule,

$$\gamma_t = 1, \tag{14}$$

which applies the same non-isotropic strength throughout the diffusion process. Following SkeletonDiffusion (Curreli et al., 2025), the *mono-decrease* schedule is defined as

$$\gamma_t = 1 - \frac{t}{T}, \tag{15}$$

which progressively weakens the non-isotropic component toward later diffusion timesteps.

We additionally explore two increasing schedules. The *beta-coupled* schedule ties the non-isotropic strength to the instantaneous diffusion noise level:

$$\gamma_t = \beta_t, \tag{16}$$

where β_t is determined by the cosine diffusion schedule. Unlike directly scheduling γ_t according to the timestep, this formulation makes the non-isotropic contribution adapt to the amount of noise introduced at each diffusion step. For comparison, we also consider a *reverse-power* schedule,

$$\gamma_t = \left(\frac{t}{T-1}\right)^p, \tag{17}$$

where $p = 1$ in our experiments. This corresponds to a linear increase from weak to strong non-isotropic corruption and is independent of the actual magnitude of β_t.

Table 4: Analysis of different non-isotropic γ_t schedules under the full model. Best and second-best results are shown in bold and underline, respectively.

γ_t Schedule	Str-Mean↓	Str-MSE↓	B1↑	B2↑	B3↑	B4↑	R↑	chrF↑
Constant	5.19	<u>18.27</u>	<u>40.17</u>	27.39	20.45	16.38	41.78	**38.95**
Mono-decrease	<u>4.98</u>	**18.10**	39.35	27.10	20.51	16.66	41.46	38.16
Beta-coupled	**4.83**	18.48	**40.55**	**27.77**	<u>20.75</u>	<u>16.68</u>	<u>41.96</u>	**38.95**
Reverse-power	5.50	18.90	40.12	<u>27.73</u>	**20.90**	**16.79**	**42.30**	<u>38.59</u>

As shown in Table 4, different γ_t schedules exhibit distinct trade-offs between structural fidelity and linguistic quality. The mono-decrease schedule achieves the lowest Str-MSE (18.10), whereas reverse-power favors higher-order linguistic metrics, obtaining the best B3 (20.90), B4 (16.79), and ROUGE (42.30), but also produces the largest stretching errors among the compared schedules.

In contrast, the beta-coupled schedule achieves the lowest Str-Mean (4.83), the best B1 (40.55) and B2 (27.77), ties for the best chrF (38.95), and remains second-best on B3, B4, and ROUGE. Unlike reverse-power, which increases the non-isotropic strength directly with the timestep regardless of the actual diffusion noise level, beta-coupling adjusts γ_t according to the instantaneous noise variance β_t. This produces a milder non-isotropic bias over low-noise steps and strengthens it as the diffusion noise increases. The results suggest that such noise-aware coupling provides a more consistent balance between skeletal structure and linguistic generation. We therefore adopt the beta-coupled schedule in our final model.

A.2 Movement Representation

We further study the choice of representative point for the movement branch. Besides the wrist joint used in our model, we consider palm-center, which represents hand motion using the center of the palm, and hand-mean, obtained by averaging the hand keypoints at each frame. As shown in Table 5, palm-center and hand-mean substantially reduce skeletal stretching, with hand-mean achieving the lowest Str-Mean and Str-MSE. However, the wrist representation consistently achieves the best back-translation performance across all linguistic metrics, including B4 (16.16) and ROUGE (42.06). This suggests that spatially averaged hand representations can produce structurally smoother motion but may suppress motion variations that are informative for sign semantics. We therefore use the wrist trajectory as the movement representation in our model.

Table 5: Analysis of movement representations. All variants use only the movement branch and differ in the representative point used to characterize hand motion. Best results are shown in bold.

Representation	Str-Mean↓	Str-MSE↓	B1↑	B2↑	B3↑	B4↑	R↑	chrF↑
Wrist	9.74	26.49	**40.70**	**27.59**	**20.38**	**16.16**	**42.06**	**38.97**
Palm-center	5.46	19.06	40.01	27.45	20.25	16.00	41.31	38.79
Hand-mean	**4.91**	**18.50**	40.53	27.28	19.99	15.76	41.12	38.49

A.3 Analysis of Movement Injection Strategy

We further analyze how the movement context is incorporated into the denoising representation under the full model. Besides the gated residual used in our method (Eq. 11), we consider a simpler alternative that directly adds the movement feature to the pose representation:

$$\mathbf{z}'_t = \mathbf{z}_t + \mathbf{f}^{(t)}_{\text{mov}}. \tag{18}$$

In contrast, Eq. 11 modulates the movement context with a learnable scalar gate before residual injection, allowing its contribution to be adaptively controlled during training.

Table 6: Comparison of movement injection strategies. Best results are shown in bold.

Injection	Str-Mean↓	Str-MSE↓	B1↑	B2↑	B3↑	B4↑	R↑	chrF↑
Full (Direct addition)	**5.08**	**18.30**	40.48	27.62	20.41	16.20	41.96	38.85
Full (Gated residual)	5.16	18.81	**40.75**	**28.22**	**21.27**	**17.11**	**42.58**	**39.20**

As shown in Table 6, direct addition achieves slightly lower Str-Mean and Str-MSE, whereas the gated residual consistently improves all linguistic metrics. In particular, B4 increases from 16.20 to 17.11 (+0.91), while ROUGE improves from 41.96 to 42.58 (+0.62). The consistent gains across BLEU, ROUGE, and chrF indicate that adaptively controlling the contribution of movement context better preserves linguistically informative motion cues, while the small difference in stretching metrics suggests comparable structural fidelity. We therefore adopt the gated residual in the final model.

A.4 Analysis of Hand Reconstruction Objective

We further analyze the reconstruction objective applied to the hand region under the full model. All variants share the same model configuration, including HS+HA non-isotropic corruption, wrist-based Movement with gated fusion, the beta-coupled γ_t schedule, and $T = 1000$ diffusion steps. The body and face regions always retain the standard MSE objective; only the reconstruction loss for the hand joints is changed. Importantly, the MSE variant in this analysis is still the full model with non-isotropic forward corruption, rather than the isotropic baseline.

Let $\mathbf{r}$ denote the reconstruction residual of the hand joints and $\mathbf{z} = \mathbf{W}_t\mathbf{r}$ its covariance-whitened counterpart. We consider the following four objectives:

$$\begin{aligned} \mathcal{L}_{\text{MSE}} &= \text{mean}(\mathbf{r}^2), \quad & \mathcal{L}_{\text{L1}} &= \text{mean}(|\mathbf{r}|), \\ \mathcal{L}_{\text{wMSE}} &= \text{mean}(\mathbf{z}^2), \quad & \mathcal{L}_{\text{wL1}} &= \text{mean}(|\mathbf{z}|). \end{aligned} \tag{19}$$

Here, $\mathbf{W}_t$ is determined by the timestep-dependent non-isotropic hand covariance. Thus, MSE and L1 operate directly in the Euclidean coordinate space, whereas wMSE and wL1 evaluate the residual after covariance-aware whitening.

Table 7: Analysis of hand reconstruction objectives under the full model. The two factors are the residual space (Euclidean or covariance-whitened) and the penalty function (squared or absolute error). Best results are shown in bold.

Residual Space	Penalty	Str-Mean↓	Str-MSE↓	B1↑	B2↑	B3↑	B4↑	R↑	chrF↑
Euclidean	MSE	9.07	28.68	40.66	27.42	20.10	15.77	42.03	38.32
Euclidean	L1	9.49	27.65	39.73	27.11	20.06	15.93	41.08	38.12
Whitened	MSE	36.92	36.92	26.21	13.31	8.89	6.68	25.22	26.75
Whitened	L1	**5.16**	**18.81**	**40.75**	**28.22**	**21.27**	**17.11**	**42.58**	**39.20**

Table 7 reveals a clear interaction between covariance whitening and the choice of reconstruction penalty. Replacing MSE with L1 directly in the Euclidean space does not improve the full model, with most linguistic metrics remaining comparable or decreasing. Likewise, applying a squared penalty after whitening performs substantially worse, indicating that covariance-aware whitening alone is insufficient when combined with a quadratic objective.

In contrast, combining covariance whitening with the L1 penalty consistently yields the strongest performance. The wL1 objective achieves the lowest Str-Mean (5.16) and improves all linguistic metrics over the Euclidean MSE variant, including B4 from 15.77 to 17.11 (+1.34), ROUGE from 42.03 to 42.58 (+0.55), and chrF from 38.32 to 39.20 (+0.88). These results suggest that the benefit arises from the interaction between structure-aware residual normalization and a robust absolute-error objective, rather than from either component alone. We therefore adopt covariance-whitened L1 reconstruction for the hand region in the full model.