Nonisotropic Gaussian Diffusion for Realistic 3D Human Motion Prediction

Supplementary Material

Contents

A. Mathematical Derivations of our Nonisotropic Gaussian Diffusion

A.1. Forward Diffusion Process

As mentioned in the main paper body, the Gaussian forward transitions are defined as:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, \boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top). \tag{14}$$

allowing us to sample from a transition in dependence of isotropic noise $\boldsymbol{\epsilon}_t$ as:

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \boldsymbol{U}\boldsymbol{\Lambda}_t^{1/2}\boldsymbol{\epsilon}_t, \tag{15}$$

We can further derive the tractable form of the forward transitions $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ by recursively applying $\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}}\boldsymbol{x}_{t-2} + \boldsymbol{U}\boldsymbol{\Lambda}_{t-1}^{1/2}\boldsymbol{\epsilon}_{t-1}$:

$$\begin{aligned}
\boldsymbol{x}_t =& \sqrt{\alpha_t}(\sqrt{\alpha_{t-1}}\boldsymbol{x}_{t-2} + \boldsymbol{U}\boldsymbol{\Lambda}_{t-1}^{1/2}\boldsymbol{\epsilon}_{t-1}) + \boldsymbol{U}\boldsymbol{\Lambda}_t^{1/2}\boldsymbol{\epsilon}_t \\
=& \sqrt{\alpha_t}(\sqrt{\alpha_{t-1}}(\sqrt{\alpha_{t-2}}\boldsymbol{x}_{t-3} \\
&+ \boldsymbol{U}\boldsymbol{\Lambda}_{t-2}^{1/2}\boldsymbol{\epsilon}_{t-2}) + \boldsymbol{U}\boldsymbol{\Lambda}_{t-1}^{1/2}\boldsymbol{\epsilon}_{t-1}) + \boldsymbol{U}\boldsymbol{\Lambda}_t^{1/2}\boldsymbol{\epsilon}_t \\
=& \dots \\
=& \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \boldsymbol{U}\bar{\boldsymbol{\Lambda}}_t^{1/2}\boldsymbol{\epsilon}_0 \\
\sim& \mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, \boldsymbol{U}(\bar{\boldsymbol{\Lambda}}_t)\boldsymbol{U}^T) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, \bar{\boldsymbol{\Sigma}}_t),
\end{aligned} \tag{16}$$

where we exploit the fact that the isotropic noises can be formulated as $\boldsymbol{\epsilon}_{t-1} \sim \mathcal{N}(\mathbf{0}, \alpha_t\boldsymbol{U}\boldsymbol{\Lambda}_{t-1}\boldsymbol{U}^T)$, $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^T)$ and that the sum of two independent Gaussian random variables is a Gaussian with mean equals the sum of the two means and the variance being the sum of the two variances. We have thus derived the Gaussian form of the tractable forward diffusion process $q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, \boldsymbol{U}\bar{\boldsymbol{\Lambda}}_t\boldsymbol{U}^T)$ for

$$\bar{\boldsymbol{\Lambda}}_t = \tilde{\gamma}_t\boldsymbol{\Lambda}_{\mathbb{I}} + (1 - \bar{\alpha}_t)\mathbb{I} \tag{17}$$

$$\begin{aligned}
\tilde{\gamma}_t =& \sum_{i=0}^{t} \bar{\gamma}_{t-i}\alpha_{t-i}^{-1} \prod_{j=t-i}^{t} \alpha_j \\
=& \bar{\alpha}_t \sum_{i=0}^{t} \frac{\bar{\gamma}_{t-i}}{\bar{\alpha}_{t-i}} = \bar{\gamma}_t + \alpha_t\tilde{\gamma}_{t-1}\,.
\end{aligned} \tag{18}$$

A.2. Reverse Diffusion Process

To perform inference, we need to find a tractable form for the posterior $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ in terms of $\boldsymbol{x}_0$. With the forms of the Gaussian transitions, through Bayes rule

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \tag{19}$$

we can start the derivation of the posterior $\mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$ from

$$\frac{\mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, \boldsymbol{\Sigma}_t)\mathcal{N}(\boldsymbol{x}_{t-1}; \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0, \bar{\boldsymbol{\Sigma}}_{t-1})}{\mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, \bar{\boldsymbol{\Sigma}}_t)}. \tag{20}$$

Differently from the conventional isotropic diffusion derivation, where this and subsequent derivations are carried out for scalar variables thanks to the i.i.d. assumption, our random variables are correlated and we have to deal with vectorial equations. Hence the posterior mean $\boldsymbol{\mu}_q$ and covariance $\boldsymbol{\Sigma}_q$ cannot be derived straightforwardly.

To address this issue, we exploit the eigenvalue decomposition of $\boldsymbol{\Sigma}_t$ and notice that the orthogonal matrix $\boldsymbol{U}$ is a linear transformation preserving the inner product of vectors by definition, and that thus the shape of the posterior probability distribution $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ stays the same in the isometry of the Euclidean space given

$$\tilde{\boldsymbol{x}}_i = \boldsymbol{U}^\top \boldsymbol{x}_i\,. \tag{21}$$

This allows us to 'rotate' the posterior distribution $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ by the transformation $\boldsymbol{U}$ and carry out the derivation for a distribution $q(\tilde{\boldsymbol{x}}_{t-1}|\tilde{\boldsymbol{x}}_t) = \mathcal{N}(\tilde{\boldsymbol{x}}_{t-1}; \tilde{\boldsymbol{\mu}}_q, \boldsymbol{\Lambda}_q)$ that now has a diagonal covariance matrix $\boldsymbol{\Lambda}_q = \boldsymbol{U}^\top\boldsymbol{\Sigma}_q\boldsymbol{U}$. Now we can handle each dimension independently, since the matrices in the following derivations are diagonal matrices and this allows us to use the commutative property $\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_2 = \boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_1$. In the following, we also make use of the observation:

$$\bar{\boldsymbol{\Lambda}}_t = \alpha_t\bar{\boldsymbol{\Lambda}}_{t-1} + \boldsymbol{\Lambda}_t \tag{22}$$

The mean and variance of the posterior can thus be derived in the isometry space as

$$\begin{aligned}
&q(\tilde{\boldsymbol{x}}_{t-1}|\tilde{\boldsymbol{x}}_t,\tilde{\boldsymbol{x}}_0) \\
&\propto \exp -\frac{1}{2}\Big[(\tilde{\boldsymbol{x}}_t - \sqrt{\alpha_t}\tilde{\boldsymbol{x}}_{t-1})^\top\boldsymbol{\Lambda}_t^{-1}(\tilde{\boldsymbol{x}}_t - \sqrt{\alpha_t}\tilde{\boldsymbol{x}}_{t-1}) \\
&\qquad +(\tilde{\boldsymbol{x}}_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\tilde{\boldsymbol{x}}_0)^\top\bar{\boldsymbol{\Lambda}}_{t-1}^{-1}(\tilde{\boldsymbol{x}}_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\tilde{\boldsymbol{x}}_0) \\
&\qquad - (\tilde{\boldsymbol{x}}_t - \sqrt{\bar{\alpha}_t}\tilde{\boldsymbol{x}}_0)^\top\bar{\boldsymbol{\Lambda}}_t^{-1}(\tilde{\boldsymbol{x}}_t - \sqrt{\bar{\alpha}_t}\tilde{\boldsymbol{x}}_0)\Big] \\
&= \exp -\frac{1}{2}\Big[\tilde{\boldsymbol{x}}_{t-1}^\top\alpha_t\boldsymbol{\Lambda}_t^{-1}\tilde{\boldsymbol{x}}_{t-1} - 2\tilde{\boldsymbol{x}}_{t-1}^\top\sqrt{\alpha_t}\boldsymbol{\Lambda}_t^{-1}\tilde{\boldsymbol{x}}_t \\
&\qquad + \tilde{\boldsymbol{x}}_{t-1}^\top\bar{\boldsymbol{\Lambda}}_{t-1}^{-1}\tilde{\boldsymbol{x}}_{t-1} - 2\tilde{\boldsymbol{x}}_{t-1}^\top\sqrt{\bar{\alpha}_{t-1}}\bar{\boldsymbol{\Lambda}}_{t-1}^{-1}\tilde{\boldsymbol{x}}_0 \\
&\qquad + C(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{x}}_0)\Big] \\
&\propto \exp -\frac{1}{2}\Big[\tilde{\boldsymbol{x}}_{t-1}^\top\underbrace{(\alpha_t\boldsymbol{\Lambda}_t^{-1} + \bar{\boldsymbol{\Lambda}}_{t-1}^{-1})}_{\boldsymbol{\Lambda}_q^{-1}}\tilde{\boldsymbol{x}}_{t-1} \\
&\qquad - 2\tilde{\boldsymbol{x}}_{t-1}^\top(\sqrt{\alpha_t}\boldsymbol{\Lambda}_t^{-1}\tilde{\boldsymbol{x}}_t + \sqrt{\bar{\alpha}_{t-1}}\bar{\boldsymbol{\Lambda}}_{t-1}^{-1}\tilde{\boldsymbol{x}}_0)\Big] \\
&= \exp -\frac{1}{2}\Big[\tilde{\boldsymbol{x}}_{t-1}^\top\boldsymbol{\Lambda}_q^{-1}\tilde{\boldsymbol{x}}_{t-1} \\
&\qquad - 2\tilde{\boldsymbol{x}}_{t-1}^\top(\sqrt{\alpha_t}\boldsymbol{\Lambda}_t^{-1}\tilde{\boldsymbol{x}}_t + \sqrt{\bar{\alpha}_{t-1}}\bar{\boldsymbol{\Lambda}}_{t-1}^{-1}\tilde{\boldsymbol{x}}_0)\Big] \\
&= \exp -\frac{1}{2}\Big[\tilde{\boldsymbol{x}}_{t-1}^\top\boldsymbol{\Lambda}_q^{-1}\tilde{\boldsymbol{x}}_{t-1} \\
&\qquad - 2\tilde{\boldsymbol{x}}_{t-1}^\top\boldsymbol{\Lambda}_q^{-1}\boldsymbol{\Lambda}_q(\sqrt{\alpha_t}\boldsymbol{\Lambda}_t^{-1}\tilde{\boldsymbol{x}}_t + \sqrt{\bar{\alpha}_{t-1}}\bar{\boldsymbol{\Lambda}}_{t-1}^{-1}\tilde{\boldsymbol{x}}_0)\Big].
\end{aligned} \tag{23}$$

Comparing Eq.(23) to $\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - 2\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + C = (\boldsymbol{x} - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$, we can describe the posterior with the following Gaussian form:

$$q(\tilde{\boldsymbol{x}}_{t-1}|\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{x}}_0) = \mathcal{N}(\tilde{\boldsymbol{x}}_{t-1}; \tilde{\boldsymbol{\mu}}_q, \boldsymbol{\Lambda}_q) \tag{24}$$

$$\begin{aligned}
\boldsymbol{\Lambda}_q &= \Big[\alpha_t\boldsymbol{\Lambda}_t^{-1} + \bar{\boldsymbol{\Lambda}}_{t-1}^{-1}\Big]^{-1} \\
&= \Big[\alpha_t\bar{\boldsymbol{\Lambda}}_{t-1}\bar{\boldsymbol{\Lambda}}_{t-1}^{-1}\boldsymbol{\Lambda}_t^{-1} + \boldsymbol{\Lambda}_t\boldsymbol{\Lambda}_t^{-1}\bar{\boldsymbol{\Lambda}}_{t-1}^{-1}\Big]^{-1} \\
&= \Big[(\alpha_t\bar{\boldsymbol{\Lambda}}_{t-1} + \boldsymbol{\Lambda}_t)\bar{\boldsymbol{\Lambda}}_{t-1}^{-1}\boldsymbol{\Lambda}_t^{-1}\Big]^{-1} \\
&= \boldsymbol{\Lambda}_t\bar{\boldsymbol{\Lambda}}_{t-1}(\boldsymbol{\Lambda}_t + \alpha_t\bar{\boldsymbol{\Lambda}}_{t-1})^{-1} \\
&\overset{(22)}{=} \boldsymbol{\Lambda}_t\bar{\boldsymbol{\Lambda}}_{t-1}\bar{\boldsymbol{\Lambda}}_t^{-1},
\end{aligned} \tag{25}$$

$$\begin{aligned}
&\tilde{\boldsymbol{\mu}}_q = \boldsymbol{\Lambda}_q(\sqrt{\alpha_t}\boldsymbol{\Lambda}_t^{-1}\tilde{\boldsymbol{x}}_t + \sqrt{\bar{\alpha}_{t-1}}\bar{\boldsymbol{\Lambda}}_{t-1}^{-1}\tilde{\boldsymbol{x}}_0) \\
&= \boldsymbol{\Lambda}_t\bar{\boldsymbol{\Lambda}}_{t-1}(\boldsymbol{\Lambda}_t + \alpha_t\bar{\boldsymbol{\Lambda}}_{t-1})^{-1}(\sqrt{\alpha_t}\boldsymbol{\Lambda}_t^{-1}\tilde{\boldsymbol{x}}_t + \sqrt{\bar{\alpha}_{t-1}}\bar{\boldsymbol{\Lambda}}_{t-1}^{-1}\tilde{\boldsymbol{x}}_0) \\
&= \bar{\boldsymbol{\Lambda}}_t^{-1}(\sqrt{\alpha_t}\bar{\boldsymbol{\Lambda}}_{t-1}\tilde{\boldsymbol{x}}_t + \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{\Lambda}_t\tilde{\boldsymbol{x}}_0)
\end{aligned} \tag{26}$$

To obtain the previous definition of $\tilde{\boldsymbol{\mu}}_q$ and $\boldsymbol{\Lambda}_q$, we make use of the following equalities, that coincide wih our intuition and understanding of denoising diffusion processes and are reported for completeness:

$$\begin{aligned}
\tilde{\gamma}_t =&\bar{\gamma}_t + \alpha_t \tilde{\gamma}_{t-1} \\
=&\bar{\gamma}_t + \alpha_t \sum_{i=0}^{t-1} \bar{\gamma}_{t-1-i} \alpha_{t-1-i}^{-1} \prod_{j=t-1-i}^{t-1} \alpha_j \\
=&\sum_{i=-1}^{-1} \bar{\gamma}_{t-1-i} \alpha_{t-1-i}^{-1} \prod_{j=t-1-i}^{t} \alpha_j \\
&+ \sum_{i=0}^{t-1} \bar{\gamma}_{t-1-i} \alpha_{t-1-i}^{-1} \prod_{j=t-1-i}^{t} \alpha_j \\
=&\sum_{i=-1}^{t-1} \bar{\gamma}_{t-1-i} \alpha_{t-1-i}^{-1} \prod_{j=t-1-i}^{t} \alpha_j \mid \text{shift the } i \text{ index by 1 } (i := i+1) \\
=&\sum_{i=0}^{t} \bar{\gamma}_{t-i} \alpha_{t-i}^{-1} \prod_{j=t-i}^{t} \alpha_j
\end{aligned} \tag{27}$$

$$\begin{aligned}
\bar{\mathbf{\Lambda}}_t =&\alpha_t \bar{\mathbf{\Lambda}}_{t-1} + \mathbf{\Lambda}_t \\
=&\alpha_t \left(\tilde{\gamma}_{t-1} \mathbf{\Lambda}_{\mathbb{I}} + (1 - \bar{\alpha}_{t-1}) \mathbb{I} \right) + \left(\bar{\gamma}_t \mathbf{\Lambda}_{\mathbb{I}} + (1 - \alpha_t) \mathbb{I} \right) \\
=&\left(\alpha_t \tilde{\gamma}_{t-1} + \bar{\gamma}_t \right) \mathbf{\Lambda}_{\mathbb{I}} + \left(\alpha_t (1 - \bar{\alpha}_{t-1}) + (1 - \alpha_t) \right) \mathbb{I} \\
=&\tilde{\gamma}_t \mathbf{\Lambda}_{\mathbb{I}} + (1 - \bar{\alpha}_t) \, \mathbb{I}
\end{aligned} \tag{28}$$

$$\bar{\mathbf{\Sigma}}_t = \alpha_t \bar{\mathbf{\Sigma}}_{t-1} + \mathbf{\Sigma}_t \tag{29}$$

We detail how to transform the new mean and covariance into the original coordinate system:

$$\begin{aligned}
\mathbf{\Sigma}_q =&\boldsymbol{U} \mathbf{\Lambda}_q \boldsymbol{U}^\top \\
=&\boldsymbol{U} \mathbf{\Lambda}_t \bar{\mathbf{\Lambda}}_{t-1} \bar{\mathbf{\Lambda}}_t^{-1} \boldsymbol{U}^\top \\
=&\boldsymbol{U} \mathbf{\Lambda}_t \underbrace{\boldsymbol{U}^\top \boldsymbol{U}}_{\mathbb{I}} \bar{\mathbf{\Lambda}}_{t-1} \boldsymbol{U}^\top \boldsymbol{U} \bar{\mathbf{\Lambda}}_t^{-1} \boldsymbol{U}^\top \\
=&\mathbf{\Sigma}_t \bar{\mathbf{\Sigma}}_{t-1} \bar{\mathbf{\Sigma}}_t^{-1},
\end{aligned} \tag{30}$$

$$\begin{aligned}
\boldsymbol{\mu}_q =&\boldsymbol{U} \tilde{\boldsymbol{\mu}}_q \\
=&\boldsymbol{U} \bar{\mathbf{\Lambda}}_t^{-1} (\sqrt{\alpha_t} \bar{\mathbf{\Lambda}}_{t-1} \tilde{\boldsymbol{x}}_t + \sqrt{\bar{\alpha}_{t-1}} \mathbf{\Lambda}_t \tilde{\boldsymbol{x}}_0) \\
=&\boldsymbol{U} \bar{\mathbf{\Lambda}}_t^{-1} \boldsymbol{U}^\top \boldsymbol{U} (\sqrt{\alpha_t} \bar{\mathbf{\Lambda}}_{t-1} \boldsymbol{U}^\top \boldsymbol{U} \tilde{\boldsymbol{x}}_t + \sqrt{\bar{\alpha}_{t-1}} \mathbf{\Lambda}_t \boldsymbol{U}^\top \boldsymbol{U} \tilde{\boldsymbol{x}}_0) \\
=&\bar{\mathbf{\Sigma}}_t^{-1} (\sqrt{\alpha_t} \bar{\mathbf{\Sigma}}_{t-1} \boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}} \mathbf{\Sigma}_t \boldsymbol{x}_0)
\end{aligned} \tag{31}$$

A.3. Training objective

Denoising diffusion probabilistic models [31] are trained by minimizing the negative log likelihood of the evidence lower bound, which can be simplified to the KL divergence between the posterior $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ and the learned reverse process $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$. Since the covariance matrix is independent of θ, the KL-divergence can be expressed as Mahalanobis distance

$$\begin{aligned}
&\arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \| p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2} \left[(\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_q)^\top \mathbf{\Sigma}_q^{-1} (\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_q) \right].
\end{aligned} \tag{32}$$

Regressing the true latent $\boldsymbol{x}_0$ We compute the KL divergence in the isometry space with diagonal covariances as

$$\begin{aligned}
&\arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(q(\tilde{\boldsymbol{x}}_{t-1}|\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{x}}_0) \| p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_{t-1}|\tilde{\boldsymbol{x}}_t)) \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2} \left[(\tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{\mu}}_q)^\top \mathbf{\Lambda}_q^{-1} (\tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{\mu}}_q) \right] \\
&= \left[\bar{\mathbf{\Lambda}}_t^{-1} \sqrt{\bar{\alpha}_{t-1}} \mathbf{\Lambda}_t (\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0) \right]^\top \mathbf{\Lambda}_q^{-1} \left[\bar{\mathbf{\Lambda}}_t^{-1} \sqrt{\bar{\alpha}_{t-1}} \mathbf{\Lambda}_t (\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0) \right] \\
&= \left[(\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0) \right]^\top \bar{\alpha}_{t-1} \mathbf{\Lambda}_q^{-1} \bar{\mathbf{\Lambda}}_t^{-2} \mathbf{\Lambda}_t^2 \left[\bar{\mathbf{\Lambda}}_t^{-1} \sqrt{\bar{\alpha}_{t-1}} \mathbf{\Lambda}_t (\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0) \right] \\
&= \left[\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0 \right]^\top \bar{\alpha}_{t-1} \mathbf{\Lambda}_t^{-1} \bar{\mathbf{\Lambda}}_{t-1}^{-1} \bar{\mathbf{\Lambda}}_t \bar{\mathbf{\Lambda}}_t^{-2} \mathbf{\Lambda}_t^2 \left[\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0 \right] \\
&= \left[\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0 \right]^\top \bar{\alpha}_{t-1} \bar{\mathbf{\Lambda}}_{t-1}^{-1} \bar{\mathbf{\Lambda}}_t^{-1} \mathbf{\Lambda}_t \left[\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0 \right] \\
&= \left[\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0 \right]^\top \bar{\mathbf{\Lambda}}_{t-1}^{-1} \bar{\mathbf{\Lambda}}_t^{-1} \bar{\alpha}_{t-1} (\bar{\mathbf{\Lambda}}_t - \alpha_t \bar{\mathbf{\Lambda}}_{t-1}) \left[\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0 \right] \\
&= \left[\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0 \right]^\top \bar{\mathbf{\Lambda}}_{t-1}^{-1} \bar{\mathbf{\Lambda}}_t^{-1} (\bar{\alpha}_{t-1} \bar{\mathbf{\Lambda}}_t - \bar{\alpha}_t \bar{\mathbf{\Lambda}}_{t-1}) \left[\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0 \right] \\
&= \left[\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0 \right]^\top (\bar{\alpha}_{t-1} \bar{\mathbf{\Lambda}}_{t-1}^{-1} - \bar{\alpha}_t \bar{\mathbf{\Lambda}}_t^{-1}) \left[\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0 \right] \\
&= \left[\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0 \right]^\top (\|\tilde{\boldsymbol{\mu}}_{t-1}\|^2 \bar{\mathbf{\Lambda}}_{t-1}^{-1} - \|\tilde{\boldsymbol{\mu}}_t\|^2 \bar{\mathbf{\Lambda}}_t^{-1}) \left[\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0 \right] \\
&= \left[\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0 \right]^\top (\mathrm{S\tilde{N}R}(t-1) - \mathrm{S\tilde{N}R}(t)) \left[\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0 \right] \\
&= \|\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0\|^2_{(\mathrm{S\tilde{N}R}(t-1) - \mathrm{S\tilde{N}R}(t))^{-1}} \\
&= \|\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0\|^2_{\boldsymbol{S}^{-1}}
\end{aligned} \tag{33}$$

where we employ the definition of $\mathrm{S\tilde{N}R}(t)) = \|\tilde{\boldsymbol{\mu}}_t\|^2 \bar{\mathbf{\Lambda}}_t^{-1}$ for the signal-to-noise ratio. The last line denotes the Mahalanobis distance between $\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{x}}_0$ with respect to a probability distribution with symmetric positive-definite covariance matrix $\boldsymbol{S} = (\mathrm{S\tilde{N}R}(t-1) - \mathrm{S\tilde{N}R}(t))^{-1}$.

As in conventional diffusion training [31], we train directly with $\boldsymbol{S} = (\mathrm{S\tilde{N}R}(t))^{-1}$, which in our case translates to $\boldsymbol{S}^{-1} = \bar{\alpha}_t \bar{\mathbf{\Lambda}}_t^{-1}$. According to the spectral theorem, for every positive-definite matrix $\boldsymbol{A}$ it holds $\boldsymbol{A}^{-1} = \boldsymbol{W}^\top \boldsymbol{W}$. Since S is diagonal, the spectral theorem translates to $\boldsymbol{S}^{-1} = \boldsymbol{S}^{-1/2\top} \boldsymbol{S}^{-1/2} = \bar{\alpha}_t \bar{\mathbf{\Lambda}}_t^{-1}$ with $\boldsymbol{W}^\top := \boldsymbol{S}^{-1/2} = \sqrt{\bar{\alpha}_t} \bar{\mathbf{\Lambda}}_t^{-1/2} = \boldsymbol{W}$ and the Mahalanobis distance becomes

$$\begin{aligned}
&\arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(q(\tilde{\boldsymbol{x}}_{t-1}|\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{x}}_0) \| p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_{t-1}|\tilde{\boldsymbol{x}}_t)) \\
&= \|\boldsymbol{W}(\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0)\|^2 = \bar{\alpha}_t \|\bar{\mathbf{\Lambda}}_t^{-1/2} (\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{x}}_0)\|^2
\end{aligned} \tag{34}$$

Thus in the original coordinate system the final training objective can be defined as

$$\begin{aligned}
&\arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \| p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) \\
&= \bar{\alpha}_t \|\bar{\mathbf{\Lambda}}_t^{-1/2} \boldsymbol{U}^\top (\boldsymbol{x}_{\boldsymbol{\theta}} - \boldsymbol{x}_0)\|^2
\end{aligned} \tag{35}$$

Regressing the noise $\boldsymbol{\epsilon_\theta}$ We report here the necessary equations for regressing the noise $\boldsymbol{\epsilon_\theta}$ instead of the true latent variable $\boldsymbol{x}_0$. By applying the reparameterization trick in the isometry space we define

$$\tilde{\boldsymbol{x}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(\tilde{\boldsymbol{x}}_t - \boldsymbol{\Lambda}_t^{1/2}\boldsymbol{\epsilon}_0) \tag{36}$$

By regressing the noise and considering the previous formulation we derive the KL-divergence with an analogous procedure.

$$\begin{aligned} &\arg\min_{\boldsymbol{\theta}} D_{\text{KL}}(q(\tilde{\boldsymbol{x}}_{t-1}|\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{x}}_0) \| p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_{t-1}|\tilde{\boldsymbol{x}}_t)) \\ &= [\boldsymbol{\epsilon}_0 - \boldsymbol{\epsilon_\theta}]^\top \frac{\boldsymbol{\Lambda}_t}{\bar{\alpha}_t}(\text{S}\tilde{\text{N}}\text{R}(t-1) - \text{S}\tilde{\text{N}}\text{R}(t))\,[\boldsymbol{\epsilon}_0 - \boldsymbol{\epsilon_\theta}] \end{aligned} \tag{37}$$

The training objective in the original covariance space is given by

$$\begin{aligned} &\arg\min_{\boldsymbol{\theta}} D_{\text{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \| p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) \\ &= [\boldsymbol{\epsilon}_0 - \boldsymbol{\epsilon_\theta}]^\top \frac{\boldsymbol{\Sigma}_t}{\bar{\alpha}_t}(\text{SNR}(t-1) - \text{SNR}(t))\,[\boldsymbol{\epsilon}_0 - \boldsymbol{\epsilon_\theta}] \end{aligned} \tag{38}$$

A.4. Alternative Nonisotropic Formulations of $\boldsymbol{\Sigma}_t$

In this section, we present formulations of the covariance of the forward noising transitions $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, \boldsymbol{\Sigma}_t)$ alternative to our nonisotropic formulation with scheduler γ_t defined in Eq. (5). We report these alternative formulations either because we ablate against them, or because these were discarded in early research stages. Note that for all formulations, the derivation of the tractable forward and posterior still holds, just for a different choice of $\bar{\boldsymbol{\Lambda}}_t$.

A.4.1 Scheduler $\gamma_t = 1$

The most straightforward case of nonisotropic Gaussian diffusion can be obtained by setting $\gamma_t = 1$ in our Eq. (5)

$$\boldsymbol{\Sigma}_t = (1-\alpha_t)\boldsymbol{\Sigma}_N = \boldsymbol{U}(1-\alpha_t)\boldsymbol{\Lambda}_N\boldsymbol{U}^\top\,, \tag{39}$$

$$\boldsymbol{\Lambda}_t = (1-\alpha_t)\boldsymbol{\Lambda}_N \tag{40}$$

resulting in nonisotropic noise sampling for the last hierarchical latent $t = T$. We highlight that this choice of $\boldsymbol{\Sigma}_t$ corresponds to performing conventional isotropic diffusion ($\boldsymbol{\Sigma}_t = \mathbb{I}$) in a normalized space where the dimensions are not correlated anymore (for example through an affine transformation disentangling the joint dimensions, or layer normalization) and transform back the diffused features to the skeleton latent space.

For the tractable form of the forward process $p(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, \boldsymbol{U}\bar{\boldsymbol{\Lambda}}_t\boldsymbol{U}^T)$ it follows

$$\bar{\boldsymbol{\Lambda}}_t = (1-\bar{\alpha}_t)\boldsymbol{\Lambda}_N \tag{41}$$

The computation of the corresponding posterior exploits the following equality:

$$\begin{aligned} \bar{\boldsymbol{\Lambda}}_t =& \alpha_t\bar{\boldsymbol{\Lambda}}_{t-1} + \boldsymbol{\Lambda}_t \\ =& \alpha_t(1-\bar{\alpha}_{t-1})\boldsymbol{\Lambda}_N + (1-\alpha_t)\boldsymbol{\Lambda}_N \\ =& (\alpha_t(1-\bar{\alpha}_{t-1}) + (1-\alpha_t))\boldsymbol{\Lambda}_N \\ =& (\alpha_t - \alpha_t\bar{\alpha}_{t-1} + 1 - \alpha_t)\boldsymbol{\Lambda}_N \\ =& (1-\bar{\alpha}_t)\boldsymbol{\Lambda}_N \end{aligned} \tag{42}$$

A.4.2 Discarded Scheduler Formulation

As a preliminary study of our correlated diffusion approach, we explored the following covariance:

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_N\alpha_t + \mathbb{I}(1-\alpha_t) \tag{43}$$

$$\boldsymbol{\Lambda}_N = \boldsymbol{\Lambda}_N\alpha_t + (1-\alpha_t)\mathbb{I} \tag{44}$$

As $\boldsymbol{\Sigma}_t \to \mathbb{I}$ for $t \to T$, we have an identity covariance matrix in the final timestep. Adding large quantities of nonisotropic noise in early diffusion timesteps as described did not yield satisfactory results during experiments. Hence this formulation was discarded at an early research stage. For completeness, we report the covariances of the tractable forward transition as

$$\bar{\boldsymbol{\Lambda}}_t = \tilde{\alpha}_t\boldsymbol{\Lambda}_N + (1-\bar{\alpha}_t)\mathbb{I} \tag{45}$$

where

$$\tilde{\alpha}_t = \sum_{i=0}^{t}\prod_{j=t-i}^{t}\alpha_j = \alpha_t(1+\tilde{\alpha}_{t-1})\,. \tag{46}$$

B. Network architecture

SkeletonDiffusion's architecture builds on top of Typed-Graph (TG) convolutions [67], a type of graph convolutions designed particularly for human motion prediction. The conditional autoencoder consists of two shallow TG GRU [67]. To obtain a strong temporal representation of arbitrary length, thus fitting both observation and ground truth future, we pass the encoder's last GRU state to a TG convolutional layer [67]. The denoiser network consists of a custom architecture of stacked residual blocks of TG convolutions and TG Attention layers. Details are available through the code implementation.

Typed Graph Attention We introduce Typed Graph Attention (TG Attention) as multi head self-attention deployed

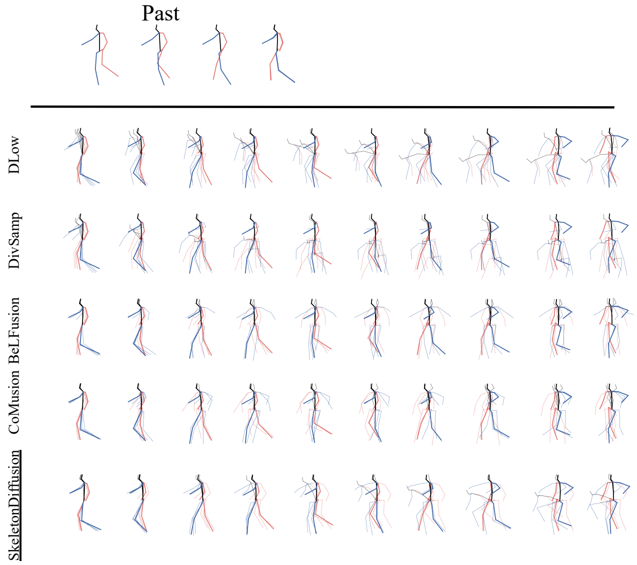

Figure 8. Qualitative Results on H36M through overlapping skeletons. Action labeled WalkTogether, segment n. 791. For each method, we display the ground truth future (thicker skeleton) overlapped by the closest prediction and the two most diverse. See Fig. 20 for a different visualization of the same qualitative.

through TG convolutions [67]. To compute scaled dot-product attention as defined by Vaswani et al. [75] with a scaling factor d_k

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}(\frac{\mathbf{QK}^T}{\sqrt{d_k}})\mathbf{V} \tag{47}$$

we define the query, key, and value matrices $\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i \in \mathbb{R}^{J\times D_{out}}$ for each head i with input $\boldsymbol{x} \in \mathbb{R}^{J\times D_{in}}$:

$$\mathbf{Q}_i = f(\text{RMS}(\boldsymbol{x})), \mathbf{K}_i = f(\text{RMS}(\boldsymbol{x})), \mathbf{V}_i = f(\text{RMS}(\boldsymbol{x})), \tag{48}$$

where f denotes the TG convolution operation described in Eq. (9) and RMS the Root Mean Square Norm (RMS)[92], acting as a regularization technique increasing the rescaling invariance of the model [75, 92].

C. Training Details

The conditional autoencoder is trained for 300 epochs on AMASS, 200 on FreeMan, and 100 on H36M. In the autoencoder training, to avoid collapse towards the motion mean of the training data [9, 80], we employ curricular learning [1, 8, 80] and learn to reconstruct sequences with random length l, sampled from a discrete uniform distribution $l \sim \mathcal{U}\{1, \tilde{F}\}$. Specifically, we increase the upper bound of the motion length $\tilde{F}$ to the original future timewindow F after the first 10 epochs with a cosine scheduler. The denoiser network is trained with $T = 10$ diffusion steps and a learning rate of 0.005 for 150 epochs. We employ a cosine scheduler [59] for α_t and implement an exponential moving average of the trained diffusion model with a decay of 0.98. Inference sampling is drawn from a DDPM sampler [31]. Both networks are trained with Adam on PyTorch. The biggest version of our model (AMASS) consists of 34M parameters and is trained on a single NVIDIA GPU A40 for 6 days. For AMASS, we measure an inference time of 471 milliseconds for a single batch on a NVIDIA GPU A40, in line with the latest DM works.

D. Details on Experiment Settings

D.1. Metrics in Stochastic HMP

First, we want to evaluate whether the generated predictions $\tilde{\mathbf{Y}} \in \mathbb{R}^{N\times F\times J\times 3}$ include the data ground truth and define *precision* metrics: the Average Distance Error (ADE) measures the Euclidean distance between the ground truth $\mathbf{Y}$ and the closest predicted sequence

$$\text{ADE}(\tilde{\mathbf{Y}}, \mathbf{Y}) = \min_n \|\tilde{\mathbf{Y}}^n - \mathbf{Y}\|_2, \tag{49}$$

while the Final Distance Error (FDE) considers only the final prediction timestep F

$$\text{FDE}(\tilde{\mathbf{Y}}, \mathbf{Y}) = \min_n \|\tilde{\mathbf{Y}}^n_F - \mathbf{Y}_F\|_2. \tag{50}$$

Because of the probabilistic nature of the task, we want to relate the predicted motions not only to a single (deterministic) ground truth but to the whole ground truth data distribution. To this end, we construct an artificial *multimodal* ground truth (MMGT) [5, 91], an ensemble of motions consisting of test data motions that share a similar last observation frame. For a sample j in the dataset defined by a past observation $\mathbf{X}$ and a ground truth future $\mathbf{Y}^j$, if the distance between the last observation frame and the last observation frame of another sample m is below a threshold δ, the future of that sample m is part of the multimodal GT for j:

$${}^{\text{MM}}\mathbf{Y}^j = \{\mathbf{Y}^m | \, m : \|\mathbf{X}^m_0 - \mathbf{X}^j_0\|_2 < \delta, \, m \neq j\} \tag{51}$$

The *multimodal* versions of the precision metrics (MMADE and MMFDE) do not consider the predicted sequence closest to the ground truth, but the one closest to the MMGT

$$\text{MMADE}(\tilde{\mathbf{Y}}, {}^{\text{MM}}\mathbf{Y}) = \min_{(i,j)\in\mathcal{M}} \|\tilde{\mathbf{Y}}^i - {}^{\text{MM}}\mathbf{Y}^j\|_2 \tag{52}$$

$$\text{MMFDE}(\tilde{\mathbf{Y}}, {}^{\text{MM}}\mathbf{Y}) = \min_{(i,j)\in\mathcal{M}} \|\tilde{\mathbf{Y}}^i_F - {}^{\text{MM}}\mathbf{Y}^j{}_F\|_2 \tag{53}$$

$$\text{with } \mathcal{M} = \{(i,j) | \, i \in [1\ldots N], \, j \in [1\ldots M]\}. \tag{54}$$

While evaluation metrics involving the MMGT may have been meaningful in the early stages of SHMP, these values should be contextualized now that methods have achieved a

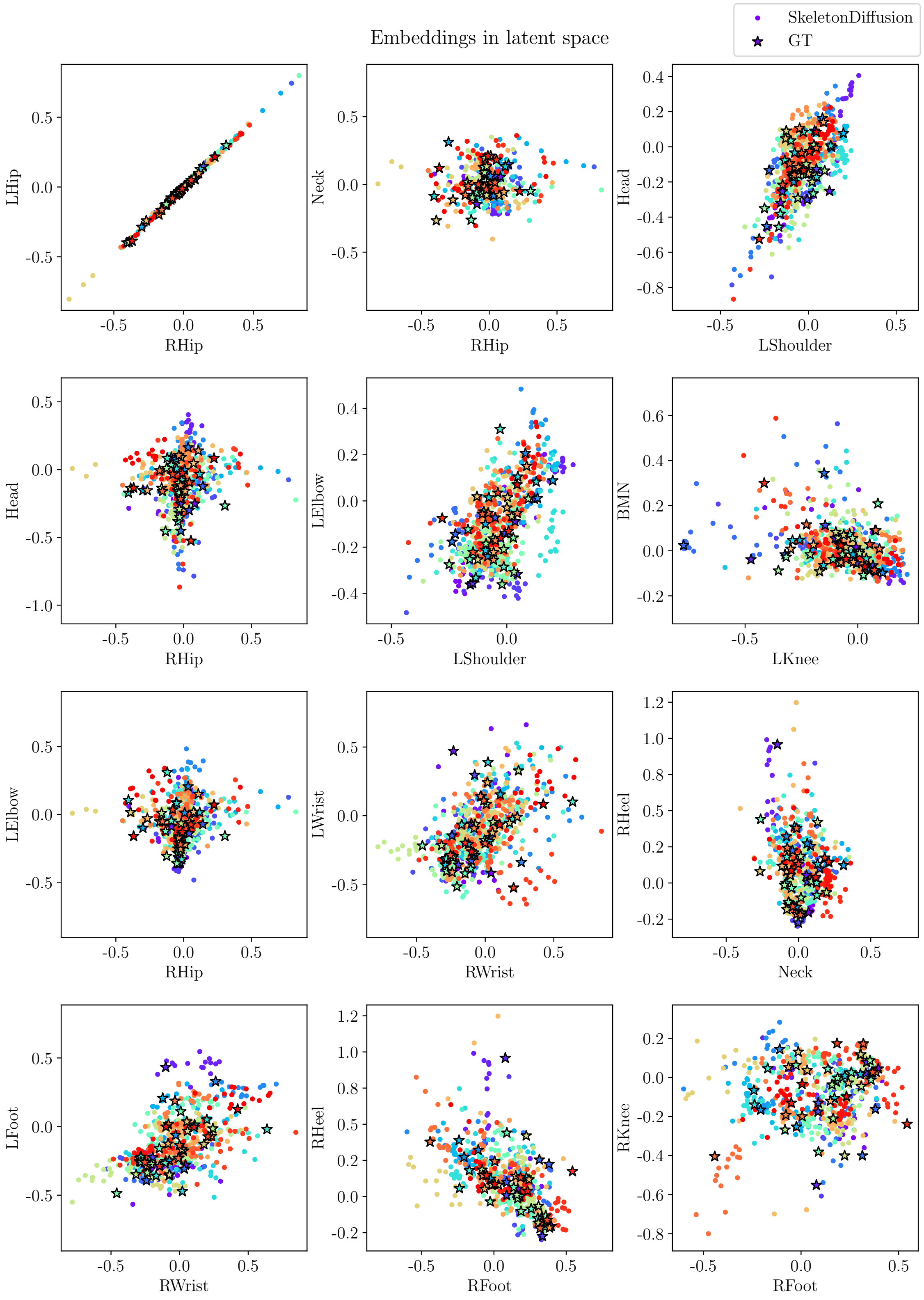

Figure 9. PCA plots of latent space embeddings for AMASS GT test segments with corresponding diffused latents generated by SkeletonDiffusion. Each GT embedding is denoted by a ⋆ of a different color, and the generated latents corresponding to the same past are denoted by a circle ○ of the same color.

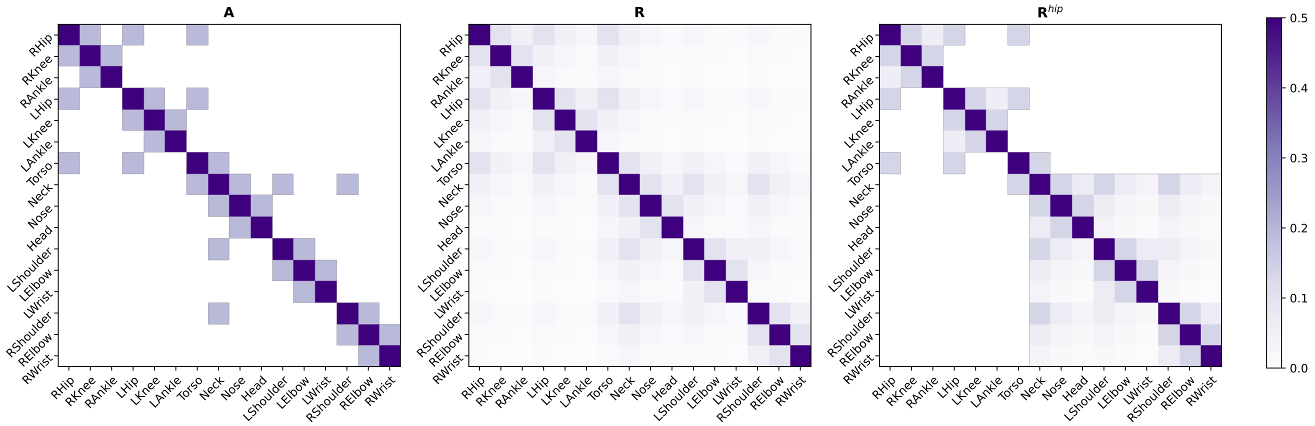

Figure 10. Node correlation matrix $\mathbf{\Sigma}_N$ for different starting choices on the H36M skeleton: the adjacency matrix $\mathbf{A}$ of the skeleton graph, the weighted transitive closure $\mathbf{R}$ and the masked weighted transitive closure $\mathbf{R}^{hip}$.

Base of $\mathbf{\Sigma}_N$	Precision		Multimodal GT		Diversity	Realism	Body Realism			
							mean ↓		RMSE ↓	
	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	APD ↑	CMD ↓	str	jit	str	jit
$\mathbf{R}$	0.481	**0.540**	0.562	**0.574**	**9.504**	11.542	3.16	0.20	4.51	0.27
$\mathbf{R}^{hip}$	**0.475**	0.543	**0.558**	0.579	8.629	12.499	**3.14**	**0.19**	**4.35**	**0.25**
$\mathbf{A}$ (SkeletonDiffusion)	0.480	0.545	0.561	0.580	9.456	**11.417**	3.15	0.20	4.45	0.26

Table 4. Ablation studies for the correlation matrix $\mathbf{\Sigma}_N$ on AMASS for adjacency matrix $\mathbf{A}$, the weighted transitive closure $\mathbf{R}$, and the masked weighted transitive closure $\mathbf{R}^{hip}$.

different level of performance: by definition, the MMGT may contain semantically inconsistent matches between past and future, which is a highly undesirable characteristic for a target distribution.

Regardless of their similarity with the ground truth data, the generated predictions should also exhibit a wide range of diverse motions. *Diversity* is measured by the Euclidean distance between motions generated from the same observation as the Average Pairwise Distance (APD):

$$\mathrm{APD}(\tilde{\mathbf{Y}}) = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \|\tilde{\mathbf{Y}}^i - \tilde{\mathbf{Y}}^j\|_2 \tag{55}$$

$$\text{with } \mathcal{P} = \{(i,j) | \, i \in [1 \dots N], \, j \in [1 \dots N], \, i \neq j\}. \tag{56}$$

Diversity can also be seen in relation to the MMGT: the Average Pairwise Distance Error (APDE) [5] measures the absolute error between the APD of the predictions and the APD of the MMGT

$$\mathrm{APDE}(\tilde{\mathbf{Y}}, {}^{\mathrm{MM}}\mathbf{Y}) = |\mathrm{APD}(\tilde{\mathbf{Y}}) - \mathrm{APD}({}^{\mathrm{MM}}\mathbf{Y})|. \tag{57}$$

Generated motions should not only be close to the GT and diverse, but also *realistic*. Barquero et al. [5] address realism in the attempt to identify speed irregularities between consecutive frames: the Cumulative Motion Distribution (CMD) measures the difference between the average joint velocity of the test data distribution $\bar{M}$ and the per-frame average velocity of the predictions M_τ.

$$\begin{aligned} \mathrm{CMD} &= \sum_{i=\tau}^{F-1} \sum_{f=1}^{\tau} \|M_\tau - \bar{M}\|_1 \\ &= \sum_{f=1}^{F-1} (F-f)\|M_\tau - \bar{M}\|_1 \end{aligned} \tag{58}$$

The Fréchet inception distance (FID) is computed for H36M only (as in [4, 13, 62]), as obtaining the necessary classifier to compute the features is not trivial: AMASS does not have class labels (recently, BABEL [62] annotated only 1% of the test data), and FreeMan's annotations do not map into specific classes.

D.2. Baselines

For the comparison on AMASS, H36M, and 3DPW we employ model checkpoints provided by the official code repositories [5, 16, 69, 82] or subsequent adaptations [5] of older models [19, 55, 79, 91]. HumanMac official repository does not provide a checkpoint for AMASS, and hence it has been discarded. For APD on H36M, MotionDiff released implementation uses a different definition which leads to significantly different results. In Tab. 9, we report the results of their checkpoint evaluated with the same metric we used for other methods.

	Precision		Multimodal GT		Diversity	Realism	Body Realism			
							mean ↓		RMSE ↓	
Norm Type	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	APD ↑	CMD ↓	str	jit	str	jit
Frob	0.480	**0.539**	0.561	**0.575**	**9.468**	12.066	3.26	0.20	4.54	0.26
Spect (SkeletonDiffusion)	0.480	0.545	0.561	0.580	9.456	**11.417**	**3.15**	**0.20**	**4.45**	**0.26**

Table 5. Ablation on the magnitude normalization procedure for $\mathbf{\Sigma}_N$ on AMASS. While normalizing with the Frobenius norm and the Spectral norm deliver very similar results, in favor of realism we opt for the spectral norm.

		Precision		Multimodal GT		Diversity	Realism
Type	param#	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	APD ↑	CMD ↓
isotropic	9M	0.509	0.571	0.576	0.598	7.875	16.229
SkeletonDiffusion		0.493	0.554	0.565	0.585	7.865	15.767
isotropic	34M	0.499	0.553	0.568	0.583	8.788	15.603
SkeletonDiffusion		0.480	0.545	0.561	0.580	9.456	11.417

Table 6. Effect of parameters number on AMASS for different types of Gaussian diffusion. Our nonisotropic diffusion training requires fewer training parameters than the isotropic formulation to reach comparable performance.

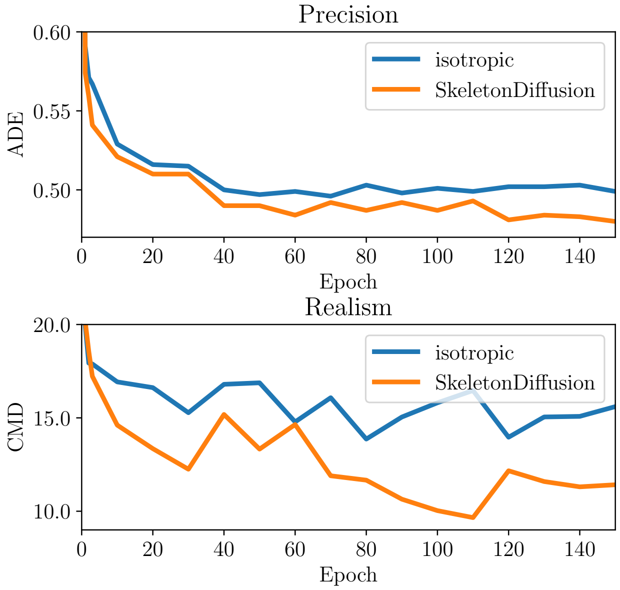

Figure 11. Our nonisotropic diffusion converges in fewer epochs than the conventional isotropic formulation.

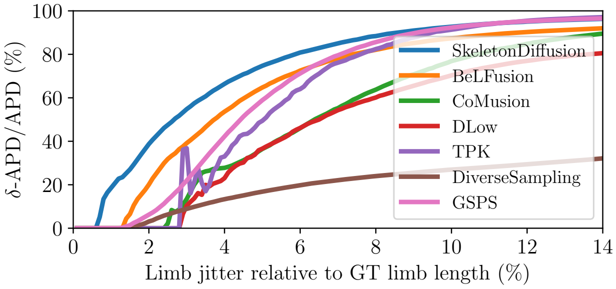

Figure 12. Diversity achieved with valid motions over total diversity according to different error tolerances on AMASS. For every method, we show the evolution of diversity (δ-APD) computed with valid motions (y-axis) for which the maximal error is below a given threshold δ (x-axis). SkeletonDiffusion presents consistently the highest diversity when considering valid poses.

D.3. Datasets

For AMASS, we follow the cross-dataset evaluation protocol proposed by Barquero et al. [5] comprising 24 datasets with a common configuration of 21 joints and a total of 9M frames with 11 datasets for training, 4 for validation, and 7 for testing with 12.7k test segments having a non-overlapping past time window. The MMGT is computed with a threshold of 0.4 resulting in an average of 125 MMGT sequences per test segment. For 3DPW, we perform zero-shot on the whole dataset merging the original splits, and by employing the same settings as AMASS we obtain 3.2k test segments with an average of 11 MMGT sequences. For H36M [34], as previous works [5, 16, 19, 19, 55, 67, 91], we train with 16 joints on subjects S1, S5, S6, S7, S8 (S8 was originally a validation subject) and test on subjects S9 and S11 with 5.2k segments for an average of 64 MMGT sequences (threshold of 0.5). FreeMan is a large-scale dataset for human pose estimation collected in-the-wild with a multi-view camera setting, depicting a wide range of actions (such as *pass ball*, *write*, *drink*, *jump rope*, and others) and 40 different actors for a total of 11M frames. As FreeMan extracts human poses from RGB, the final data may be noisy and contain ill-posed sequences. We prune the data to obtain fully labeled poses with a limb stretching lower than 5cm, and by applying the same evaluation settings as H36M obtain 11.0k test segments with an average of 69 MMGT. In the next paragraph, we report the pruning protocol. Note that as FreeMan is collected in the wild, it provides video information that could be potentially used as valuable context information for the human motion predic-

	Precision		Multimodal GT		Diversity	Realism	Body Realism			
							mean ↓		RMSE ↓	
	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	APD ↑	CMD ↓	str	jit	str	jit
w/o-TG-Att	0.502	0.567	0.576	0.597	8.021	14.934	3.90	0.20	5.31	0.27
iso	0.499	0.553	0.568	0.583	8.788	15.603	3.72	**0.18**	4.93	**0.24**
noniso	<u>0.489</u>	<u>0.547</u>	<u>0.567</u>	<u>0.581</u>	9.483	<u>11.812</u>	**2.77**	0.20	**4.06**	0.27
Ours (SkeletonDiffusion)	**0.480**	**0.545**	**0.562**	**0.579**	9.456	**11.418**	<u>3.15</u>	<u>0.20</u>	<u>4.45</u>	<u>0.26</u>

Table 7. Ablations on the AMASS dataset [53].

	Precision		Multimodal GT		Diversity	Realism	Body Realism			
							mean ↓		RMSE ↓	
	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	APD ↑	CMD ↓	str	jit	str	jit
Ours+Past	0.574	0.584	0.607	0.599	<u>9.856</u>	16.993	10.16	0.24	11.04	0.38
Ours+DCT	0.534	0.572	0.595	0.600	**11.215**	16.783	5.20	0.25	7.59	0.35
Ours (SkeletonDiffusion)	**0.480**	**0.545**	**0.562**	**0.579**	9.456	**11.418**	<u>3.15</u>	<u>0.20</u>	<u>4.45</u>	<u>0.26</u>

Table 8. Additional ablations on AMASS [53] for discarded components.

tion task for future works.

Pruning Noisy Data on FreeMan The authors of FreeMan [81] compute 3D keypoints according to different protocols, and we prefer to take the most precise data when available (*smoothnet32* over *smoothnet* over *optim* derivation). The protocols exhibit a restricted number of failure cases (for example, sudden moves very close to camera lenses). To avoid training and evaluating on strong failure cases, we remove all sequences where the difference in limb length between consecutive frames in the ground truth exceeds 5cm - a good trade-off between the overall accuracy error range of the dataset and the precision required for the task. In comparison, the maximal limb length error between consecutive frames in H36M (MoCap data) is 0.026 mm. Overall we obtain 1M frames, more than three times as much as H36M. To balance the splits after pruning, we move test subjects 1, 37, 14, 2, 12 and validation subjects 24, 18, 21 to the train split. We train on 724k densely sampled training segments (3.3k segments for validation). H36M, instead, is composed by 305k samples.

D.4. Visualization of Generated Motions.

As mentioned in the main paper, often metrics hide or may be influenced by artifacts. Inspecting qualitative results can lead to better insights into the effective SHMP methods' performance. Previous works [5, 16, 19, 55, 69, 79, 91] visualize the diversity of the predictions by overlapping the skeleton of multiple motions in different colors. This representation is limited and not well suited to identify motion irregularities. We propose to fit a SMPL mesh to each skeleton pose to ease inspection of the results, while preserving the semanticity of the prediction. Ill-posed predictions can thus be easily spotted through the erroneous SMPL fitting. For completeness, we still report the historical visualizations in Fig. 8.

E. Further Analysis

E.1. Correlations of Latent Space

We visualize the latent space in terms of the correlation among different latent joint dimensions. To this end, we embed all AMASS test segments in the latent space, and compute the first principal component along the each joint dimension separately. For each embedding, we then plot the principal component of two joint dimensions against each other. In Fig. 9, we show 50 random test segments and for each 15 diffused latents. Our latent space reflects correlations connected body joints that are expected (e.g. LHip and RHip) or are less intuitive (e.g. Neck and Hip always show in the same space direction), while other joints do not exhibit univocal correlations (e.g. Wrist and Ankle of the same body side). Weak correlations (probably related to the walking pattern) can be observed between opposite joints of the lower and upper body such as RHip and LElbow.

E.2. Discussion on Correlation Matrix $\mathbf{\Sigma}_N$

On the Magnitude Normalization The magnitude of $\mathbf{\Sigma}_N$ is constrained as in Eq. (4), where, after adding entries along the diagonal, we divide by the highest eigenvalue (spectral norm). In Tab. 5, we show results on AMASS for another normalization choice, the Frobenius norm i.e. the average of the eigenvalues. While both norms deliver very similar results, we opt for the spectral norm as the realism metrics indicate lower limb stretching and joint velocity closer to the GT data (CMD). An educated guess for the subtle difference is that higher noise magnitude (Frobenius norm) eases the generation of more diverse samples (higher diversity) but at the same time loses details of fine-grained joint positions (lower realism and limb stretching).

		Precision			Multimodal GT		Diversity	Realism		Body Realism			
										mean ↓		RMSE ↓	
Type	Method	ADE ↓	FDE ↓	MAE ↓	MMADE ↓	MMFDE ↓	APD ↑	CMD ↓	FID	str	jit	str	jit
Alg	Zero-Velocity	0.597	0.884	6.753	0.683	0.909	0.000	22.812	0.606	0.00	0.00	0.00	0.00
GAN	HP-GAN [6]	0.858	0.867	-	0.847	0.858	7.214	-	-	-	-	-	-
	DeLiGAN [30]	0.483	0.534	-	0.520	0.545	6.509	-	-	-	-	-	-
VAE	TPK [79]	0.461	0.560	8.056	0.522	0.569	6.723	6.326	0.538	6.69	0.24	8.37	0.31
	DLow [91]	0.425	0.518	6.856	0.495	0.531	11.741	4.927	1.255	7.67	0.28	9.71	0.36
	GSPS [55]	0.389	0.496	7.171	0.476	0.525	14.757	10.758	2.103	4.83	0.19	6.17	0.24
	Motron [67]	0.375	0.488	-	0.509	0.539	7.168	40.796	13.743	-	-	-	-
	DivSamp [19]	0.370	0.485	6.257	0.475	0.516	15.310	11.692	2.083	6.16	0.23	7.85	0.29
Other	STARS [84]	0.358	0.445	-	0.442	0.471	**15.884**	-	-	-	-	-	-
	SLD [83]	0.348	**0.436**	-	**0.435**	**0.463**	8.741	-	-	-	-	-	-
DM	MotionDiff [82]	0.411	0.509	-	0.508	0.536	7.254	-	-	8.04	0.59	10.21	0.77
	HumanMAC [16]	0.369	0.480	6.167	0.509	0.545	6.301	-	-	4.01	0.46	6.04	0.57
	BeLFusion [5]	0.372	0.474	6.107	0.473	0.507	7.602	5.988	0.209	5.39	0.17	6.63	0.22
	CoMusion [69]	0.350	0.458	5.904	0.494	0.506	7.632	**3.202**	**0.102**	4.61	0.41	5.97	0.56
DM	SkeletonDiff	**0.344**	0.450	5.556	0.487	0.512	7.249	4.178	0.123	**3.90**	**0.16**	**4.96**	**0.21**

Table 9. Comparison on Human3.6M [34]. Bold and underlined results correspond to the best and second-best results, respectively.

Sophistications on the Choice of $\mathbf{\Sigma}_N$ For the correlation matrix $\mathbf{\Sigma}_N$ from Eq. (4), we opt for the most straightforward and simple starting choice, the adjacency matrix $\mathbf{A}$. Here we report further studies to two more sophisticated initial choices: the weighted transitive closure $\mathbf{R}$ and the masked weighted transitive closure $\mathbf{R}^{hip}$. Given two nodes v_i and v_j in the graph, the shortest path is denoted by $P(i,j)$. The number of hops between v_i and v_j is denoted by $h_{i,j}$. We then can express the weighted transitive closure $\mathbf{R}$ as:

$$\mathbf{R}_{i,j} := \eta^{h_{i,j}-1} \tag{59}$$

with some $\eta \in (0,1)$, representing the reachability of each node weighted by the hops. As the hip joint is critical in human motion, we also consider a masked version $\mathbf{R}^{hip}$:

$$\mathbf{R}^{hip}_{i,j} = \begin{cases} \mathbf{R}_{i,j} & \text{if } v_{hip} \in P(i,j), v_i \neq v_{hip}, v_j \neq v_{hip} \\ 0 & \text{otherwise} \end{cases} \tag{60}$$

These three node correlation matrices are visualized on the H36M dataset in Fig. 10. While all three alternatives obtain good results on AMASS in Tab. 4, we opt for the adjacency matrix $\mathbf{A}$ as it is not handcrafted and allows our nonisotropic approach to generalize in a straightforward manner to different datasets. We see the analysis of sophisticated choices for $\mathbf{\Sigma}_N$ as an exciting future direction.

E.3. On the Convergence of Nonisotropic Diffusion

As depicted in Fig. 11, our nonisotropic formulation converges faster than the isotropic counterparty. As the time required for a train iteration is equal among both formulations up to a few negligible matrix multiplications, our nonisotropic formulation achieves higher performance in fewer iterations. In Tab. 6, we show that for similar performance (precision ADE) our nonisotropic formulation requires fewer parameters than conventional isotropic diffusion. We report these findings as they may be relevant for HMP applications or other structured tasks employing diffusion models.

E.4. Ablations of SkeletonDiffusion

In Tab. 7, we report the ablations discussed in Sec. 5.2. We compare the effect of TG-Attention layers on isotropic diffusion ($\mathbf{\Sigma}_N = I$ and $\gamma_t = 0$) and analyze nonisotropic diffusion with our covariance reflecting joint connections $\mathbf{\Sigma}_N$ (Eq. (4)) in the variant where $\gamma_t = 1$ (as in Eq. (3)) and our blending with the scheduler γ_t (Eq. (5)). Further experiments, such as fine-tuning the encoder responsible for embedding the past observation or representing motion data via the Discrete Cosine Transform (DCT) [16] are reported in Tab. 8. From the low precision results of the latter experiment and referring to Tab. 1, we speculate that while DCT seems suitable for transformer-based diffusion models operating in input space [16, 69], extracting features directly from Euclidean motion space seems a better choice for latent diffusion models (BeLFusion [5] and our method).

F. Additional Experiments

F.1. Diversity and Body Realism

In the main paper we discuss our intuition on how artifacts in the generated motions may lead to increased distance between the predictions and so to a better diversity metric (APD). We wish to provide evidence of this phenomenon with an argument similar to the one employed in

		Precision			Multimodal GT		Diversity	Realism	Body Realism			
									mean ↓		RMSE ↓	
Type	Method	ADE ↓	FDE ↓	MAE ↓	MMADE ↓	MMFDE ↓	APD ↑	CMD ↓	str	jit	str	jit
Alg	ZeroVelocity	0.764	1.016	10.921	0.785	1.019	0.000	40.695	4.52	0.00	4.52	0.00
VAE	DLow	0.596	0.652	9.188	0.615	0.654	13.776	12.754	8.79	0.43	11.73	0.63
	DivSamp	0.583	0.690	10.758	0.617	0.698	**23.878**	46.594	12.38	0.82	18.11	1.07
DM	BeLFusion	**0.507**	0.596	9.914	**0.543**	0.606	7.750	16.812	9.07	0.23	10.65	0.31
	CoMusion	0.550	0.600	8.773	0.588	0.611	14.400	12.282	6.21	0.66	8.60	0.87
	Ours	0.517	**0.587**	**7.106**	0.567	**0.603**	10.547	**8.188**	**4.56**	**0.22**	**5.95**	**0.30**

Table 10. Models trained on AMASS tested on zero-shot on 3DPW with synthetic noise up to 2cm added to 25% of the input.

	Precision			Body Realism			
				mean ↓		RMSE ↓	
	ADE ↓	FDE ↓	MAE ↓	str	jit	str	jit
DLow	0.716	0.776	12.397	7.36	0.23	9.57	0.40
DivSamp	0.728	0.879	12.373	5.01	0.23	7.49	0.32
BeLFusion	**0.657**	**0.756**	11.175	8.89	0.18	10.69	0.27
CoMusion	0.670	0.792	10.215	4.56	0.33	6.28	0.46
SkeletonDiffusion	0.660	0.779	**9.045**	**3.67**	**0.14**	**4.94**	**0.24**

Table 11. Long term prediction (5s) on AMASS via autoregression of models traimed to predict 2s. MMGT is undefined in this case.

	Memory↓	NumParams↓	Time↓
DLow	31 MB	8.1 M	111 ms
DivSamp	88 MB	23.1 M	8 ms
BeLFusion	53 MB	17.8 M	10 341 ms
HumanMAC	114 MB	28.7 M	7 438 ms
Comusion	87 MB	19 M	153 ms
SkeletonDiffusion	106 MB	26.5 M	412 ms

Table 12. Model footprint for a single H36M inference (RTX 6000)

Fig. 7 of the main paper i.e. by inspecting the evolution of the APD metric at different tolerance thresholds of limb jitter. First, we compute the valid motions among the generated predictions per method on the AMASS dataset, discarding a sequence if it displays a bone length jitter above a given threshold δ. By calculating the average pairwise distance APD only between valid motions and relating this value to the customary APD, in Fig. 12 we can see the contribution of ill-posed motions on diversity. Such evolving diversity differs significantly from the values reported in Tab. 1. Our method generates by a large margin the most diverse motions when considering realism according to limb jitter, demonstrating excellence also under strict constraints. Non-smooth curve regions display the influence of ill-posed motions on diversity when considering a small ensemble of predictions, as for CoMusion and TPK. When the number of valid motions is small and some of them present stretching, removing the unrealistic motions may considerably improve or worsen the average pairwise distance, resulting in sudden jumps in the curves. We are thus the first to demonstrate quantitatively that unrealistic motions increase diversity.

F.2. Human3.6M

In Tab. 9, we report quantitative results on H36M. The H36M dataset is particularly small and contains only 7 subjects. We consider this dataset less informative about generalization capabilities of the methods, and more vulnerable to overfitting. With analogous considerations as on AMASS, SkeletonDiffusion achieves state-of-the-art performance. Thanks to the explicit bias on the human skeleton, SkeletonDiffusion consistently achieves the best body realism, in particular in regard to limb stretching. Even in a setting with limited data, the prior on the skeleton structure contributes to achieving consistent realism.

Overall, the body realism metrics for DM methods appear improved compared to AMASS (Tab. 1). Along VAE and DM approaches, another line of work relies on representation learning and vocabulary techniques [83, 84]. While these methods achieve good performance, they employ carefully handcrafted loss functions, limiting the angles and bones between body joints or leveraging the multimodal ground truth in loss computations. Inconveniently, they are required to scrape the whole training data to compute the reference values or the multimodal ground truth, with computational expenses that scale quadratically with the number of instances in the dataset and require considerable engineering effort to be adapted to big data.

F.3. Challenging Scenario: Synthetic Noise in Zero-Shot Generalization

We perform further experiments on the out-of-distribution, in-the-wild data of 3D Poses in the wild (3DPW), evaluated in Tab. 3, by designing a challenging scenario with synthetic noise (Tab. 10). We add random noise of a maximal magnitude of 2cm to 25% of the input observation keypoint, thus testing robustness to noise for models that were trained with precise, MoCap data (AMASS). While the experiments in Tab. 2 show models trained on noisy data (FreeMan), here we test robustness to noise in a zero-shot setting. SkeletonDiffusiondelivers among the highest precision and diversity, and the most realistic motions with a gap between 26% and 65% compared to the otherwise closest competitor, CoMusion (see Tab. 1). While BeLFusion shows jitter values close to ours, the limb stretching and the CMD are almost double as high, meaning that the length of their limbs highly varies over the whole prediction timespan, and the joint velocities are unrealistic: they achieve high precision with extremely unrealistic motions.

F.4. Long Term Prediction

We test models trained on AMASS to predict the next 2s in the generation of 5s motions via autoregression (Tab. 11). Here we focus on Precision and Realism, as the multimodal GT is ill-defined in this setting, and diversity evaluation loses meaning and its measuremnt is polluted by the difficulty of the task. We achieve again the highest realism and SoTA precision demonstrating the effectivness of our explicit bias on the human skeleton.

F.5. Computational Efficiency

Measurement are reported in Tab. 12. While there is no obvious computational difference between diffusion models in latent (BeLFusion, Ours) and input space (HumanMAC, CoMusion), latent models achieve much better body realism, particularly jitter (Tab. 1), by not working with 3D coordinates directly.

G. More Qualitative Examples

We show more qualitative results on AMASS in Figs. 13, 14, 15, 16 and 17. More qualitative examples for H36M can be found in Figs. 18, 19 and 20 and Fig. 8.

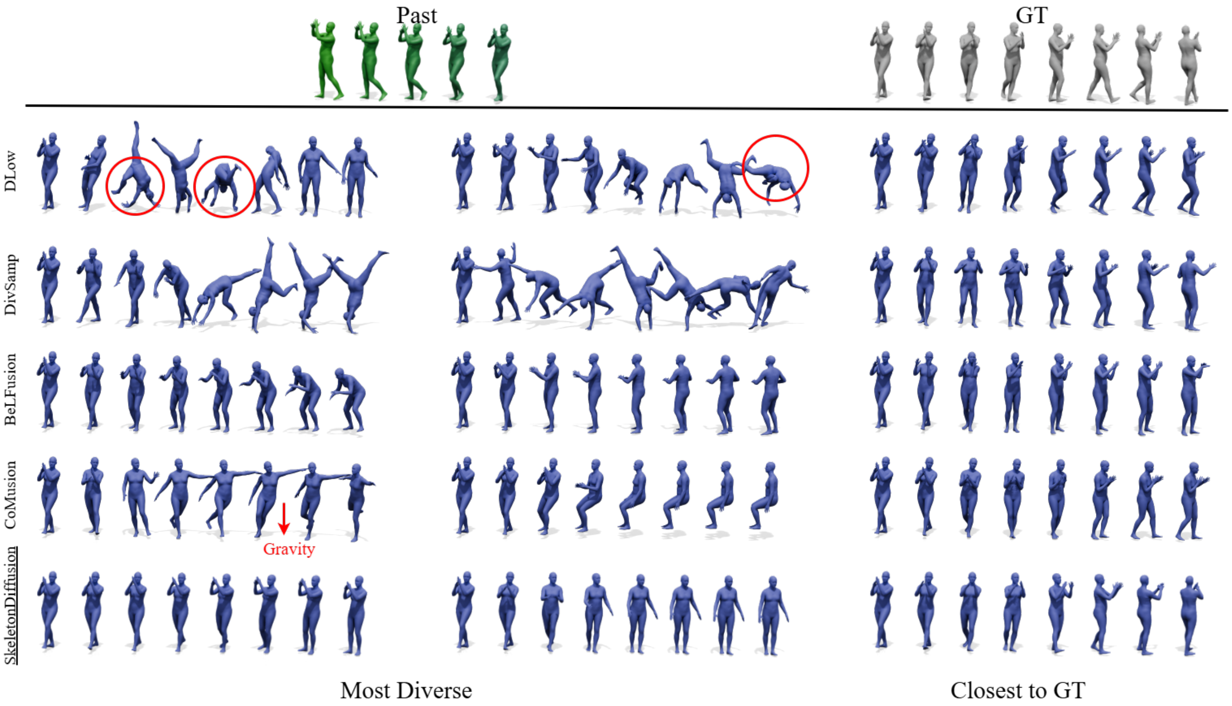

Figure 13. Qualitative Results on AMASS. From DanceDB dataset, segment n. 4122.

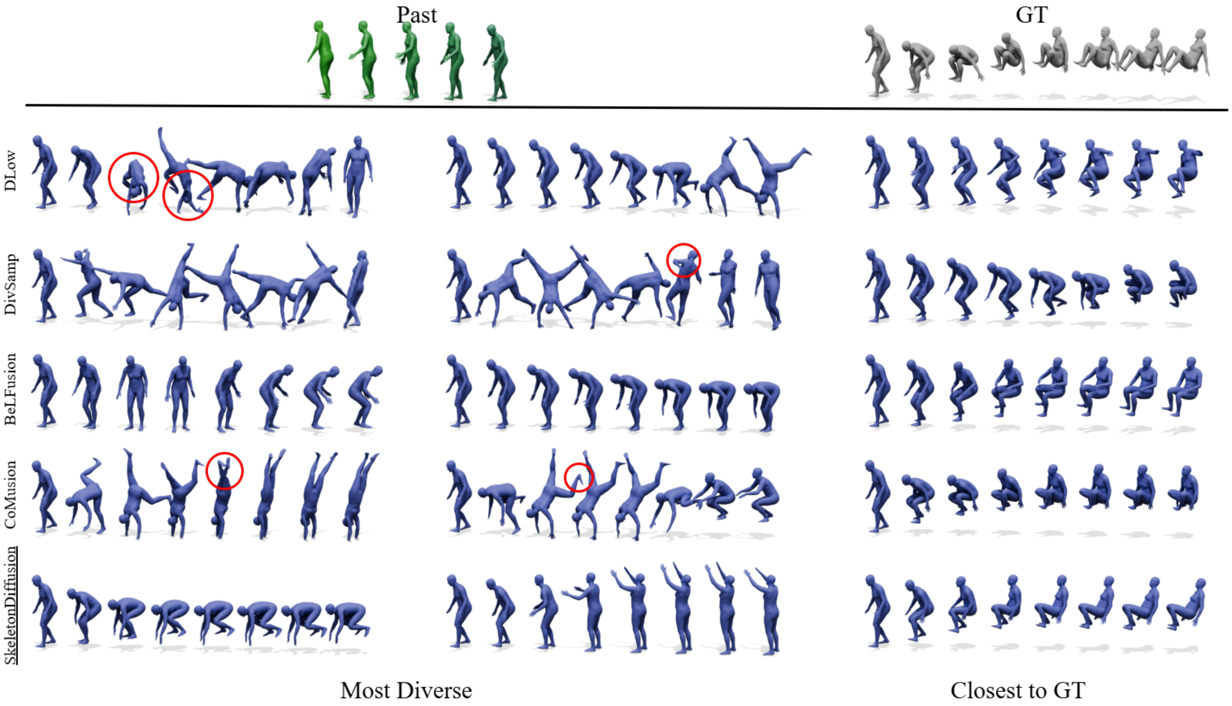

Figure 14. Qualitative Results on AMASS. From Human4D dataset, segment n. 11949.

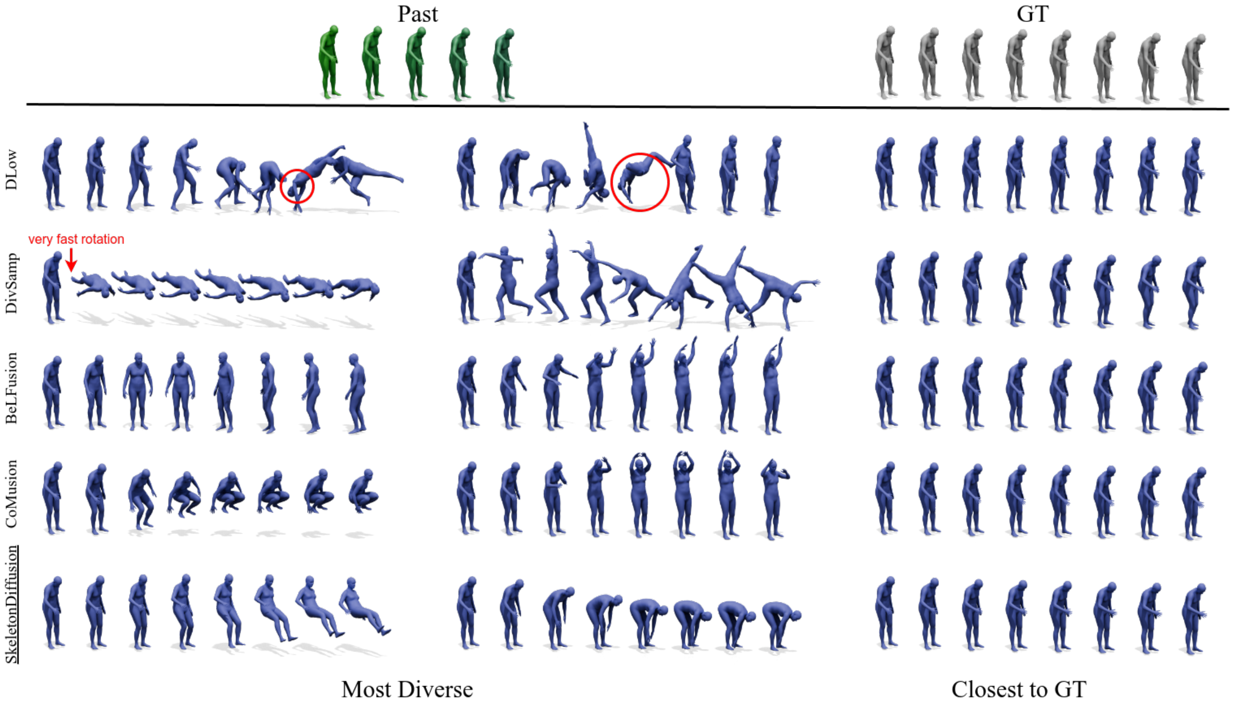

Figure 15. Qualitative Results on AMASS. From GRAB dataset, segment n. 9622.

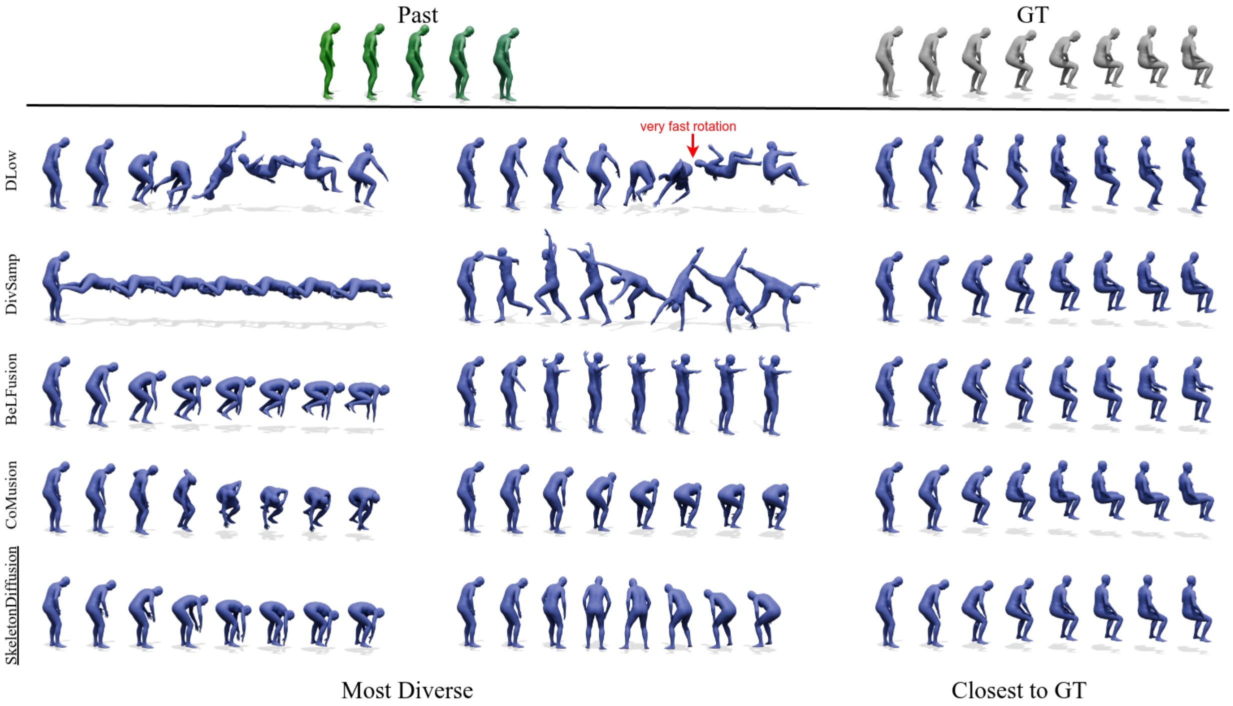

Figure 16. Qualitative Results on AMASS. From Human4D dataset, segment n. 12267.

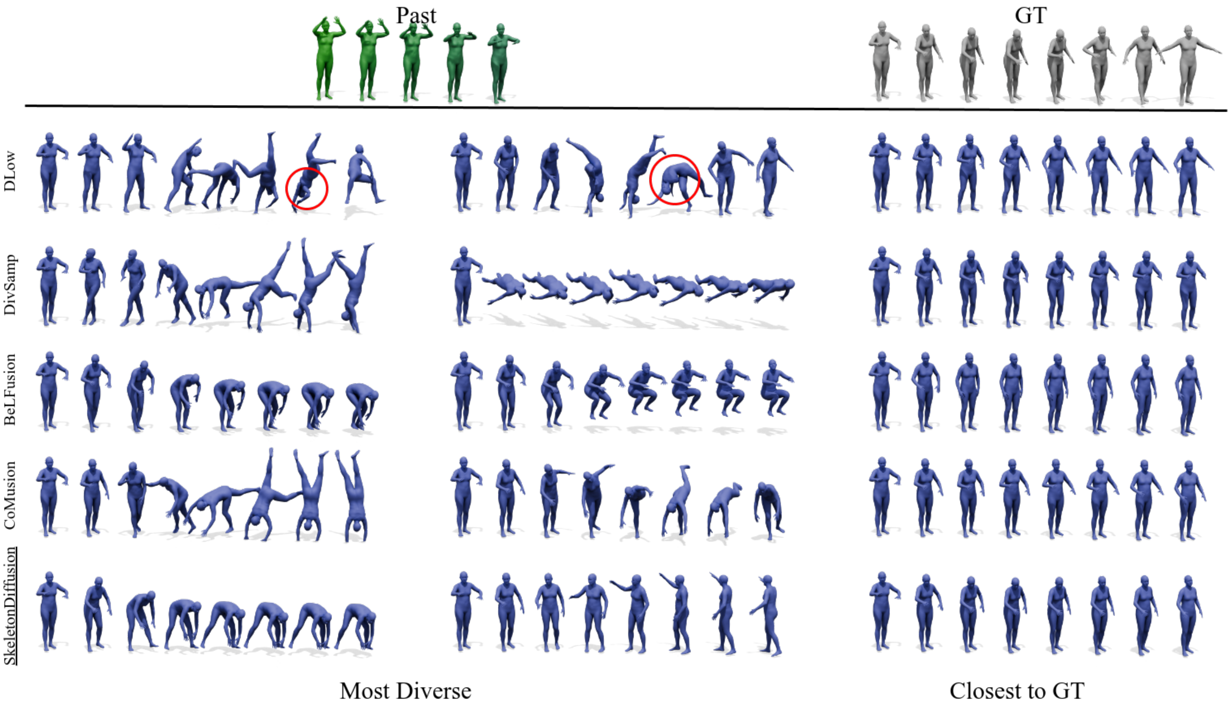

Figure 17. Qualitative Results on AMASS. From GRAB dataset, segment n. 10188.

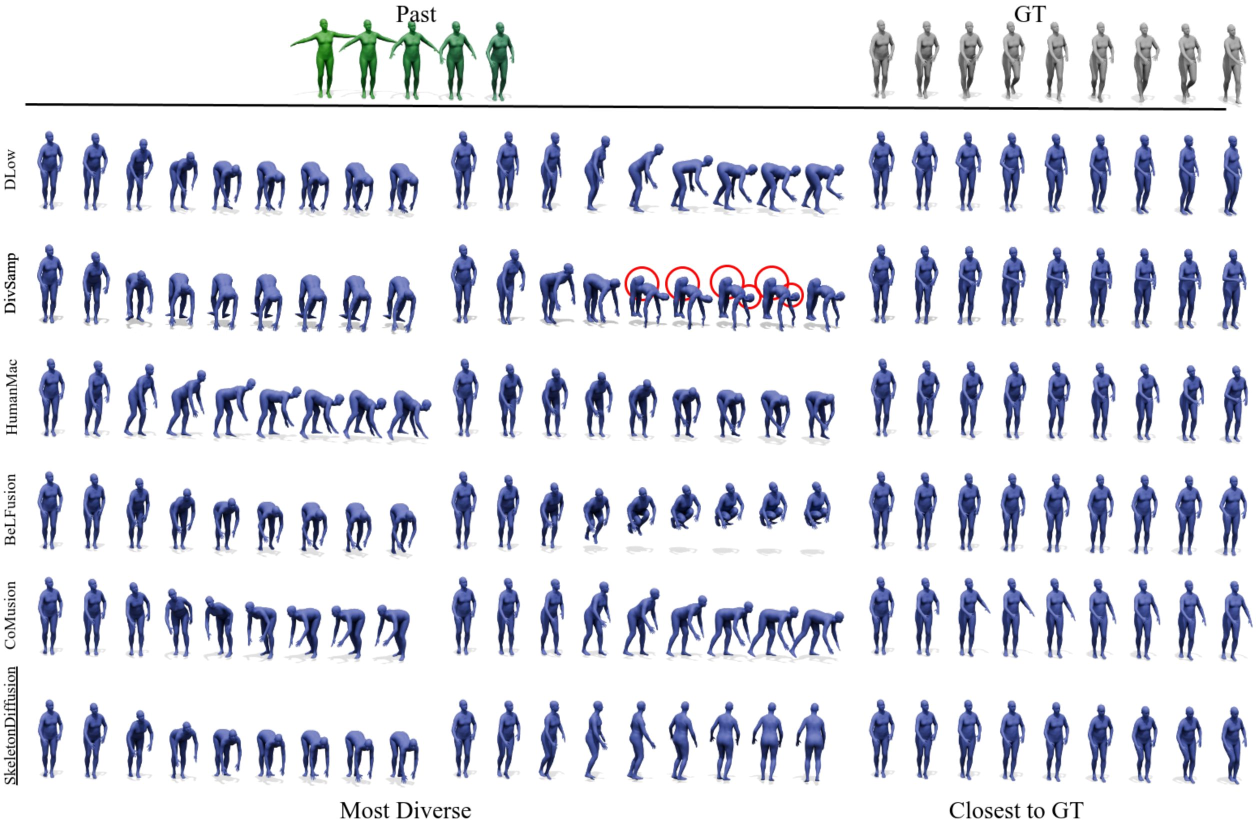

Figure 18. Qualitative Results on H36M. Action labeled WalkDog, segment n. 3122.

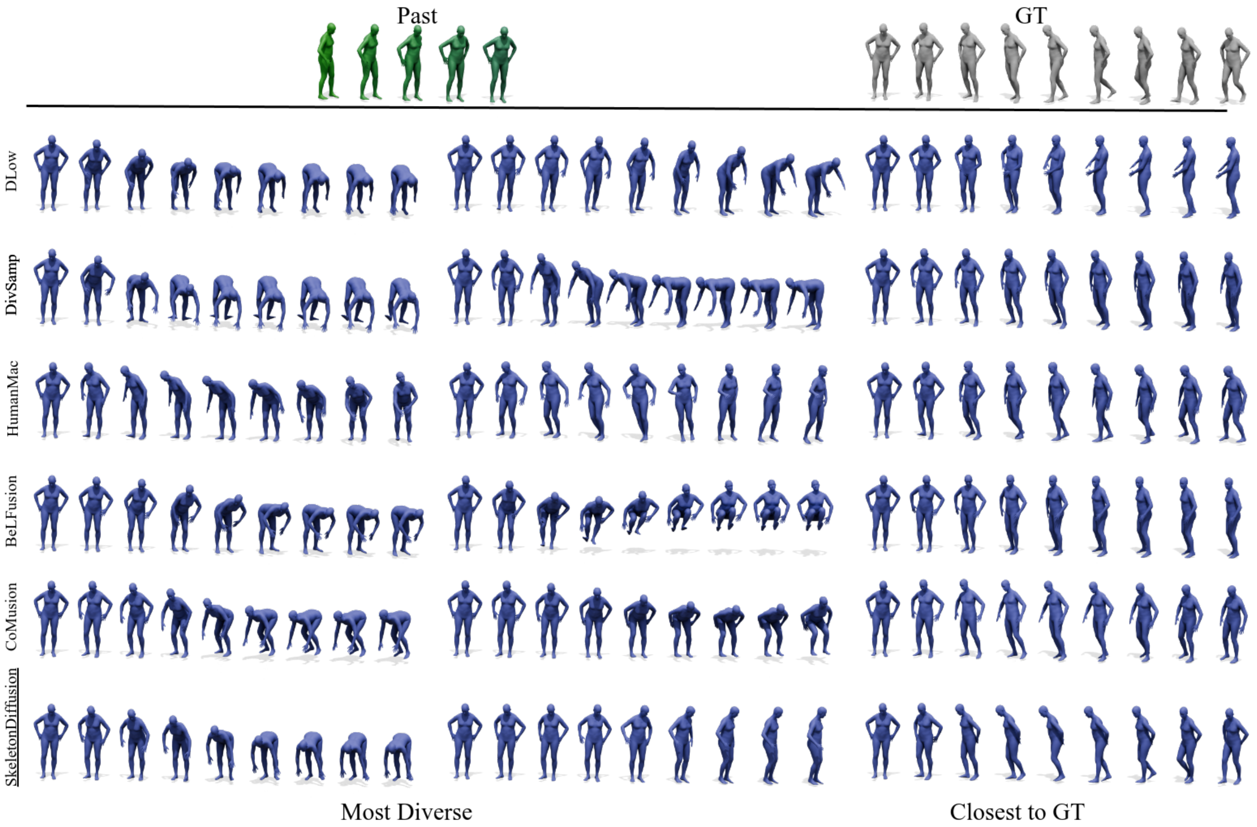

Figure 19. Qualitative Results on H36M. Action labeled Discussion, segment n. 2620.

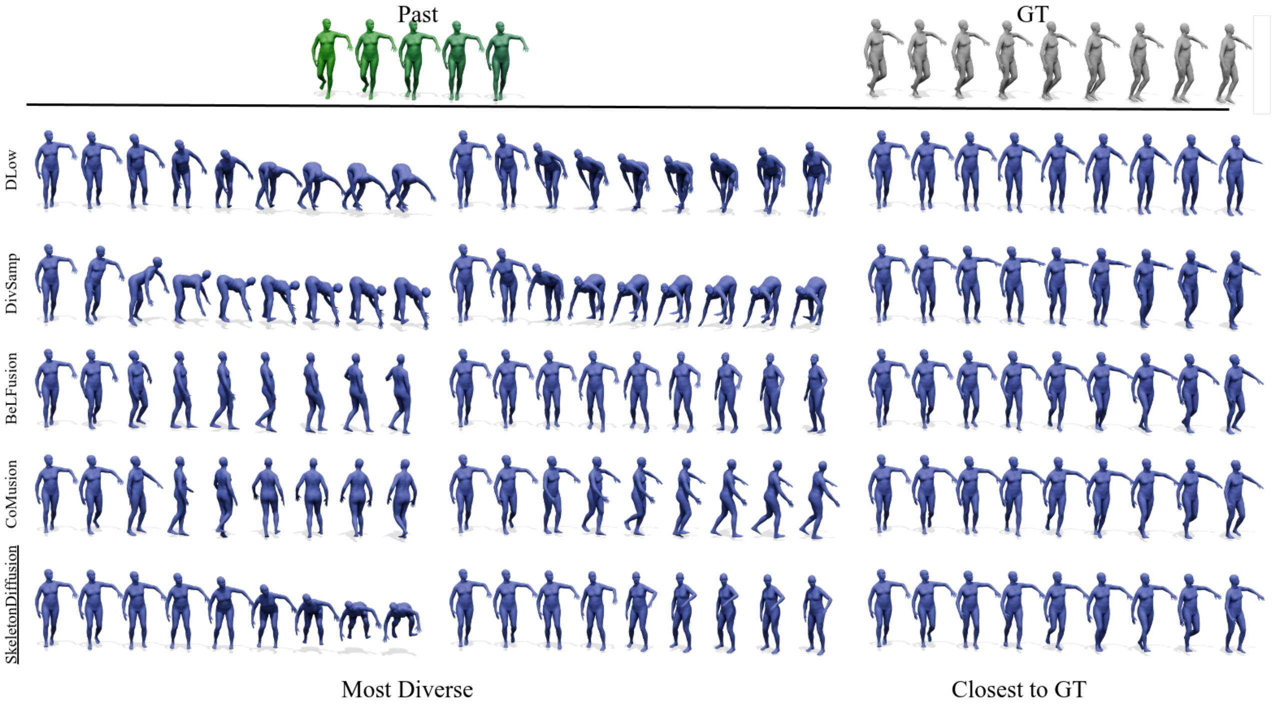

Figure 20. Qualitative Results on H36M. Action labeled WalkTogether, segment n. 791.