Supplementary Materials

S1 Marginal distribution under uniform direction prior

The von Mises-Fisher-Gaussian distribution is a compound distribution

$$u \sim \text{vMF}(\nu, \kappa)$$
$$x \mid u \sim \mathcal{N}(\mu + \rho u, \sigma^2 I),$$

with the probability density function given by (Mukhopadhyay et al., 2019)

$$f_{\text{vMFG}}(x \mid \mu, \sigma, \rho, \nu, \kappa) = (2\pi\sigma^2)^{-d/2} \frac{C_d(\kappa)}{C_d(\|\kappa\nu + \rho(x-\mu)/\sigma^2\|_2)} \exp\Big\{ -\frac{1}{2\sigma^2} \left(\|x-\mu\|_2^2 + \rho^2 \right) \Big\}, \tag{1}$$

where $C_d(\kappa) = (2\pi)^{-d/2} \frac{\kappa^{d/2-1}}{I_{d/2-1}(\kappa)}$ is the normalizing constant of a vMF distribution of dimension d with concentration κ, and I_p denotes the modified Bessel function of the first kind and order p and is defined as

$$I_p(\kappa) = \left(\frac{\kappa}{2}\right)^p \frac{1}{\Gamma\left(p + \frac{1}{2}\right) \Gamma\left(\frac{1}{2}\right)} \int_{-1}^{1} e^{\kappa t} (1-t^2)^{p-\frac{1}{2}} \, \mathrm{d}t.$$

When $d = 3$, as in this work, the modified Bessel function and the normalizing constant simplify to

$$I_{1/2}(\kappa) = \frac{(\kappa/2)^{1/2}}{\Gamma(1)\Gamma\left(\frac{1}{2}\right)} \int_{-1}^{1} e^{\kappa t} \, \mathrm{d}t = \left(\frac{\kappa}{2\pi}\right)^{1/2} \frac{1}{\kappa} \left(e^{\kappa} - e^{-\kappa}\right) = (2\pi\kappa)^{-1/2} \left(e^{\kappa} - e^{-\kappa}\right)$$
$$C_3(\kappa) = (2\pi)^{-3/2} \frac{\kappa^{1/2}}{I_{1/2}(\kappa)} = \frac{\kappa}{2\pi\left(e^{\kappa} - e^{-\kappa}\right)} = \frac{\kappa}{4\pi \sinh(\kappa)}.$$

The marginal distribution of x is a continuous mixture of Gaussians with means on the surface of a sphere centered at μ and radius ρ. Samples of x are thus concentrated around the surface of the sphere with variance set by σ^2 and direction governed by ν and κ.

A special special case of the vMFG distribution arises when $\kappa = 0$, yielding a distribution on distances $r := \|x-\mu\|_2$. The vMF distribution reduces to the uniform distribution on the sphere, and the vMFG distribution becomes only a function of μ, σ, and ρ,

$$\begin{aligned}
f_{\text{vMFG}}(x \mid \mu, \sigma, \rho, \nu, \kappa = 0) &= (2\pi\sigma^2)^{-3/2} \frac{C_3(0)}{C_3\left(\frac{\rho\|x-\mu\|_2}{\sigma^2}\right)} \exp\Big\{ -\frac{1}{2\sigma^2} \left(\|x-\mu\|_2^2 + \rho^2 \right) \Big\} \\
&= \frac{(2\pi\sigma^2)^{-3/2}}{4\pi \cdot C_3\left(\frac{\rho\|x-\mu\|_2}{\sigma^2}\right)} \exp\Big\{ -\frac{1}{2\sigma^2} \left(\|x-\mu\|_2^2 + \rho^2 \right) \Big\} \\
&= \frac{(2\pi\sigma^2)^{-1}}{4\pi \cdot C_3\left(\frac{\rho\|x-\mu\|_2}{\sigma^2}\right)} e^{-\frac{\rho\|x-\mu\|_2}{\sigma^2}} \mathcal{N}(\|x-\mu\|_2 \mid \rho, \sigma^2) \\
&= \frac{1}{4\pi\rho r} \left(1 - e^{-2\frac{\rho\|x-\mu\|_2}{\sigma^2}}\right) \mathcal{N}(\|x-\mu\|_2 \mid \rho, \sigma^2)
\end{aligned}$$

where $C_3(0) = (4\pi)^{-1}$ because $\frac{\kappa}{\sinh(\kappa)} \to 1$ as $\kappa \to 0$.

Since this is a spherically symmetric distribution on $x \in \mathbb{R}^3$, we can derive the distribution on distances $r = \|x - \mu\|_2 \in \mathbb{R}_+$ by integrating over the spherical shell of points with thickness $\mathrm{d}r$ and distance r:

$$\begin{aligned} f(r \mid \sigma, \rho)\, \mathrm{d}r &= f_{\mathrm{vMFG}}(x \mid \|x - \mu\|_2 = r; \mu, \sigma, \rho, \nu, \kappa = 0) \cdot 4\pi r^2\, \mathrm{d}r \\ &= \frac{1}{4\pi\rho r}\left(1 - e^{-2\frac{\rho r}{\sigma^2}}\right) \mathcal{N}(r \mid \rho, \sigma^2) \cdot 4\pi r^2\, \mathrm{d}r \\ &= \frac{r}{\rho}\left(1 - e^{-2\frac{\rho r}{\sigma^2}}\right) \mathcal{N}(r \mid \rho, \sigma^2)\, \mathrm{d}r. \end{aligned} \tag{2}$$

This function describes a proper distribution on distances r. For example, note that $f(r = 0) \mid \cdot) = 0$ always holds true, as would be desired of such a distribution. When variance is small, relative to mean distance, this results in high concentration of samples about the mean. Distances from the origin are then distributed as a normal distribution (i.e. first row of Figure S1). As variance increases, however, the normal distribution becomes increasingly inappropriate (i.e. subsequent rows of Figure S1), because high variance samples result in a positive bias away from the mean distance. These results are demonstrated empirically in Figure S1. In each case, the vMFG distribution under the uniform directional prior (eq. (2)) exactly fits the empirical histogram.

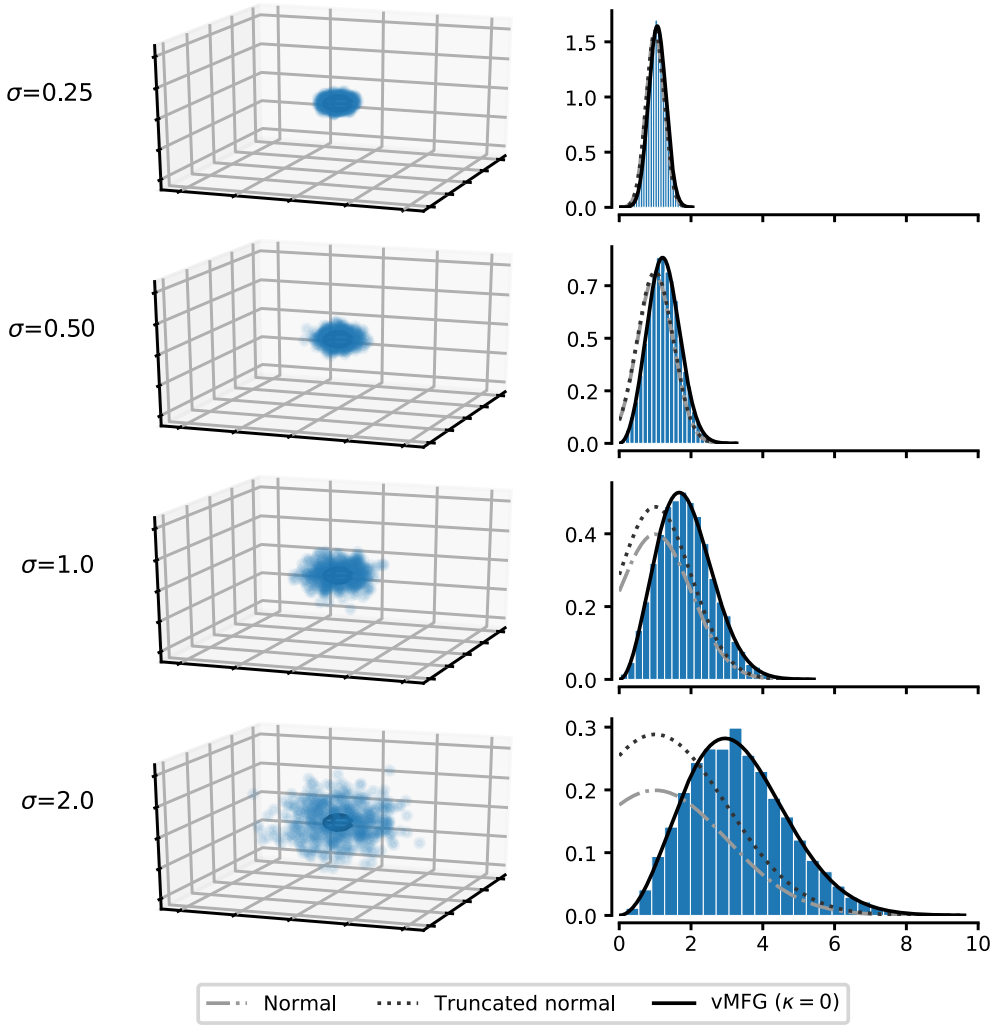

Figure S1: Left. Scatterplot of random position samples that lie $\rho = 1$ from the origin and with varying standard deviation σ. Black wireframe sphere with radius 1 is plotted for reference. *Right.* Histogram of distance of sampled positions from the origin (25 bins), versus probability density functions of the normal distribution (light grey, dashed-dotted line) evaluated for $r \geq 0$, the truncated normal distribution with lower bound 0 (darker grey, dotted line), and the vMFG distribution under a uniform direction prior (black, solid line).

S2 Conditional distribution of hierarchical von Mises-Fisher-Gaussian model

We strip away temporal considerations and observation models to consider just the hierarchical von Mises-Fisher-Gaussian distribution of our model. Recall that we define a tree-structured graph $\mathcal{G} = \{(\pi(k), k)\}_{k=2}^{K}$ where the keypoints are ordered such that keypoint 1 is the root node and each subsequent keypoint $k > 1$ has one parent $\pi(k) \in \{1, \ldots, k-1\}$. The marginal distribution of each keypoint $x_k \in \mathbb{R}^3$, given direction vector from parent $u_k \in \mathbb{S}_2$, for $k = 2, \ldots, K$ is generated by

$$u_k \sim \text{vMF}(\nu_k, \kappa_k) \tag{3}$$

$$x_k \mid u_k, x_{\pi(k)} \sim \mathcal{N}(x_{\pi(k)} + \rho_k u_{t,k}, \sigma^2 I), \tag{4}$$

for fixed parameters length $\rho_k > 0$ on edge $(\pi(k), k)$, variance σ^2, mean direction ν_k, and concentration κ_k. The marginal distribution of the root node is given by

$$x_1 \sim \mathcal{N}(\mu_1, \sigma_{k,1}^2 I), \tag{5}$$

with hyperparameters root location μ_1 and root variance σ_1. In the animal pose estimation task, set $\sigma_1 \ll \sigma_k$ if the root node position μ_1 is known *a priori* and fixed (e.g. subject is restrained); set $\sigma_1 \gg \sigma_k$ if the root location is to be inferred (e.g. subject is allowed to behave freely).

Given u_k, the mean of x_k is linear in the position of its parent. Therefore, the keypoint positions $\mathbf{x} = \{x_k\}_{k=1^K}$ are jointly Gaussian distributed when conditioned on $\mathbf{u} = \{u_k\}_{k=2^K}$,

$$p(\boldsymbol{x} \mid \boldsymbol{u}) \sim \mathcal{N}(\mu, \Sigma). \tag{6}$$

In the following, we derive expressions for the parameters of this distribution.

$$\begin{aligned}
p(\boldsymbol{x} \,|\, \boldsymbol{u}) &= \mathcal{N}(x_1 \,|\, \mu_1, \sigma_1^2 I) \prod_{k=2}^{K} \mathcal{N}(x_k \,|\, x_{\pi(k)} + \rho_k u_k, \sigma_k^2 I) \\
&\propto \exp\Big\{ -\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^\mathsf{T}(x_1 - \mu_1) - \sum_{k=2}^{K} \frac{1}{2\sigma_k^2}\big(x_k - (x_{\pi(k)} + \rho_k u_k)\big)^\mathsf{T}\big(x_k - (x_{\pi(k)} + \rho_k u_k)\big)\Big\} \\
&\propto \exp\Bigg\{ -\frac{1}{2}\Big(\frac{\mu_1^\mathsf{T}\mu_1}{\sigma_1^2} + \sum_{k=2}^{K} \frac{\rho_k^2 u_k^\mathsf{T} u_k}{\sigma_k^2}\Big) + \Big(\frac{\mu_1^\mathsf{T}}{\sigma_1^2} x_1 + \sum_{k=2}^{K} \Big(\frac{\rho_k u_k^\mathsf{T}}{\sigma_k^2} x_k - \frac{\rho_k u_k^\mathsf{T}}{\sigma_k^2} x_{\pi(k)} \Big) \Big) \\
&\qquad\qquad - \frac{1}{2}\Big(\sum_{k=1}^{K} \Big(\frac{x_k^\mathsf{T} x_k}{\sigma_k^2} \Big) + \sum_{k=2}^{K} \Big(\frac{x_{\pi(k)}^\mathsf{T} x_{\pi(k)}}{\sigma_{\pi(k)}^2} \Big) - \frac{x_{\pi(k)}^\mathsf{T} x_k}{\sigma_k^2} \Big) \Big\}.
\end{aligned} \tag{7}$$

Aside from the trouble of careful accounting of interacting terms between keypoints and their parents, the conditional expression contains a quadratic term in u_k. However, since we define $u_k \in \mathbb{S}_2$, $u_k^\mathsf{T} u_k = 1$. Therefore, there remain no u_k's and the term simplifies

$$\sum_{k=2}^{K} \frac{\rho_k^2 u_k^\mathsf{T} u_k}{\sigma_k^2} = \sum_{k=2}^{K} \frac{\rho_k^2}{\sigma_k^2},$$

and we are left with a distribution that is simply Gaussian distributed. The mean parameters of this distribution are $\Sigma = L^{-1} \in \mathbb{R}^{3K \times 3K}$ and $\mu = L^{-1} h \in \mathbb{R}^{3K}$, where

$$h_1 = \frac{\mu_1}{\sigma_1^2} - \sum_{\ell : \pi(\ell) = k} \frac{\rho_\ell}{\sigma_\ell^2} u_\ell \tag{8}$$

$$h_k = \frac{\rho_k}{\sigma_k^2} u_k - \sum_{\ell : \pi(\ell) = k} \frac{\rho_\ell}{\sigma_\ell^2} u_\ell \qquad \text{for } k = 2, \ldots, K \tag{9}$$

$$L_{jk} = \begin{cases} \sigma_k^{-2} + \sum_{\ell : \pi(\ell) = k} \sigma_\ell^{-2} & \text{for } j = k \\ -\sigma_k^{-2} & \text{for } j = \pi(k) \end{cases} \qquad \text{for } k = 1, \ldots, K. \tag{10}$$

Note that the precision matrix L takes the form of a graph Laplacian matrix with 3×3 blcoks, with diagonal elements weighted by the sum of the covariances of the keypoint and its children, and off-diagonal elements weight by the child covariance.

S3 Conditional distribution of heading angles

Here, we derive the conditional distribution of h_t given $\mathbf{u}_t, s_t$ under a uniform prior. Since there is no temporal component, we omit the subscript t in the following. Our generative model is

$$h \sim \mathrm{vM}(0, 0)$$
$$\boldsymbol{u} \mid h \sim \prod_j^J \mathrm{vMF}(\mathrm{R}(h)\,\mu_j, \kappa_j)$$

where $h \in [-\pi, \pi]$, $u \in \mathbb{S}^2$. Note that we parametrize the von Mises (vM) with a trignometric (angular) parameter and the von Mises-Fisher (vMF) distributions with unit vector parameters. $\mathrm{R}(\cdot)$ is the 3-dimensional rotation matrix that performs a rotation in the xy-plane. We will show that the posterior of h is also distributed according to a von Mises distribution

$$\begin{aligned} p(h \mid \boldsymbol{u}) &\propto \mathrm{vM}(0, 0) \cdot \prod_k^J \mathrm{vMF}(\mathrm{R}(h)\,\mu_k, \kappa_k) \\ &\propto \prod_k^K \exp\left\{\kappa_k \left(\mathrm{R}(h)\,\mu_k\right)^\mathsf{T} u_k\right\} \\ &\propto \frac{1}{2\pi I_0(\tau)} \exp\{\tau \cos(\theta - h)\} = \mathrm{vM}(\theta, \tau). \end{aligned}$$

We will make heavy use of cosine angle sum identity in this derivation,

$$\cos(\alpha \pm \beta) = \cos(\alpha)\cos(\beta) \mp \sin(\alpha)\sin(\beta).$$

We begin by writing our unit vectors μ_k and u_k in their trigonometric forms. Let $\sphericalangle_{\mathsf{xz}}(v)$ and $\sphericalangle_{\mathsf{xy}}(v)$ denote the azimuthal and polar angles, respectively, of a unit vector $v \in \mathbb{S}_2$. Then,

$$\mu_k = \begin{bmatrix} \sin(\sphericalangle_{\mathsf{xz}}(\mu_k))\,\cos(\sphericalangle_{\mathsf{xy}}(\mu_k)) \\ \sin(\sphericalangle_{\mathsf{xz}}(\mu_k))\,\sin(\sphericalangle_{\mathsf{xy}}(\mu_k)) \\ \cos(\sphericalangle_{\mathsf{xz}}(\mu_k)) \end{bmatrix}, \qquad u_k = \begin{bmatrix} \sin(\sphericalangle_{\mathsf{xz}}(u_k))\,\cos(\sphericalangle_{\mathsf{xy}}(\mu_k)) \\ \sin(\sphericalangle_{\mathsf{xz}}(u_k))\,\sin(\sphericalangle_{\mathsf{xy}}(\mu_k)) \\ \cos(\sphericalangle_{\mathsf{xz}}(u_k)) \end{bmatrix}.$$

For a single joint, we calculate

$$\begin{aligned} (\mathrm{R}(h)\,\mu_k)^\mathsf{T} u_k &= \left(\begin{bmatrix} \cos(h) & -\sin(h) & 0 \\ \sin(h) & \cos(h) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \sin(\phi_j)\,\cos(\sphericalangle_{\mathsf{xy}}(\mu_k)) \\ \sin(\sphericalangle_{\mathsf{xz}}(\mu_k))\,\sin(\sphericalangle_{\mathsf{xy}}(\mu_k)) \\ \cos(\sphericalangle_{\mathsf{xz}}(\mu_k)) \end{bmatrix} \right)^\mathsf{T} \begin{bmatrix} \sin(\sphericalangle_{\mathsf{xz}}(u_k))\,\cos(\sphericalangle_{\mathsf{xy}}(u_k)) \\ \sin(\sphericalangle_{\mathsf{xz}}(u_k))\,\sin(\sphericalangle_{\mathsf{xy}}(u_k)) \\ \cos(\sphericalangle_{\mathsf{xz}}(u_k)) \end{bmatrix} \\ &= \cos(h)\,\sin(\sphericalangle_{\mathsf{xz}}(\mu_k))\,\cos(\sphericalangle_{\mathsf{xy}}(\mu_k))\,\sin(\sphericalangle_{\mathsf{xz}}(u_k))\,\cos(\sphericalangle_{\mathsf{xy}}(u_k)) \\ &\quad - \sin(h)\,\sin(\sphericalangle_{\mathsf{xz}}(\mu_k))\,\sin(\sphericalangle_{\mathsf{xy}}(\mu_k))\,\sin(\sphericalangle_{\mathsf{xz}}(u_k))\,\cos(\sphericalangle_{\mathsf{xy}}(u_k)) \\ &\quad + \sin(h)\,\sin(\sphericalangle_{\mathsf{xz}}(\mu_k))\,\cos(\sphericalangle_{\mathsf{xy}}(\mu_k))\,\sin(\sphericalangle_{\mathsf{xz}}(u_k))\,\sin(\sphericalangle_{\mathsf{xy}}(u_k)) \\ &\quad + \cos(h)\,\sin(\sphericalangle_{\mathsf{xz}}(\mu_k))\,\sin(\sphericalangle_{\mathsf{xy}}(\mu_k))\,\sin(\sphericalangle_{\mathsf{xz}}(u_k))\,\sin(\sphericalangle_{\mathsf{xy}}(u_k)) \\ &\quad + \cos(\sphericalangle_{\mathsf{xz}}(\mu_k))\,\cos(\sphericalangle_{\mathsf{xz}}(u_k)) \\ &= \sin(\sphericalangle_{\mathsf{xz}}(\mu_k))\,\sin(\sphericalangle_{\mathsf{xz}}(u_k))\,\Big(\cos(h)\,\big(\cos(\sphericalangle_{\mathsf{xy}}(\mu_k))\,\cos(\sphericalangle_{\mathsf{xy}}(u_k)) + \sin(\sphericalangle_{\mathsf{xy}}(\mu_k))\,\sin(\sphericalangle_{\mathsf{xy}}(u_k))\big) \\ &\quad + \sin(h)\,\big(\cos(\sphericalangle_{\mathsf{xy}}(\mu_k))\,\sin(\sphericalangle_{\mathsf{xy}}(u_k)) - \sin(\sphericalangle_{\mathsf{xy}}(\mu_k))\,\cos(\sphericalangle_{\mathsf{xy}}(u_k))\big)\Big) + \cos(\sphericalangle_{\mathsf{xz}}(\mu_k))\,\cos(\sphericalangle_{\mathsf{xz}}(u_k)) \\ &= \sin(\sphericalangle_{\mathsf{xz}}(\mu_k))\,\sin(\sphericalangle_{\mathsf{xz}}(u_k))\,\Big(\cos(h)\,\cos(\sphericalangle_{\mathsf{xy}}(\mu_k) - \sphericalangle_{\mathsf{xy}}(u_k)) + \sin(h)\,\sin(\sphericalangle_{\mathsf{xy}}(u_k) - \sphericalangle_{\mathsf{xy}}(\mu_k))\Big) \\ &\quad + \cos(\sphericalangle_{\mathsf{xz}}(\mu_k))\,\cos(\sphericalangle_{\mathsf{xz}}(u_k)) \\ &= \sin(\sphericalangle_{\mathsf{xz}}(\mu_k))\,\sin(\sphericalangle_{\mathsf{xz}}(u_k))\,\Big(\cos(h)\,\cos(\sphericalangle_{\mathsf{xy}}(u_k) - \sphericalangle_{\mathsf{xy}}(\mu_k)) + \sin(h)\,\sin(\sphericalangle_{\mathsf{xy}}(u_k) - \sphericalangle_{\mathsf{xy}}(\mu_k))\Big) \\ &\quad + \cos(\sphericalangle_{\mathsf{xz}}(\mu_k))\,\cos(\sphericalangle_{\mathsf{xz}}(u_k)) \\ &= \sin(\sphericalangle_{\mathsf{xz}}(\mu_k))\,\sin(\sphericalangle_{\mathsf{xz}}(u_k))\,\cos(h - (\sphericalangle_{\mathsf{xy}}(u_k) - \sphericalangle_{\mathsf{xy}}(\mu_k))) + \cos(\sphericalangle_{\mathsf{xz}}(\mu_k))\,\cos(\sphericalangle_{\mathsf{xz}}(u_k)) \\ &= \sin(\sphericalangle_{\mathsf{xz}}(\mu_k))\,\sin(\sphericalangle_{\mathsf{xz}}(u_k))\cos(h - (\sphericalangle_{\mathsf{xy}}(u_k) - \sphericalangle_{\mathsf{xy}}(\mu_k))) + \cos(\sphericalangle_{\mathsf{xz}}(\mu_k))\,\cos(\sphericalangle_{\mathsf{xz}}(u_k)). \end{aligned}$$

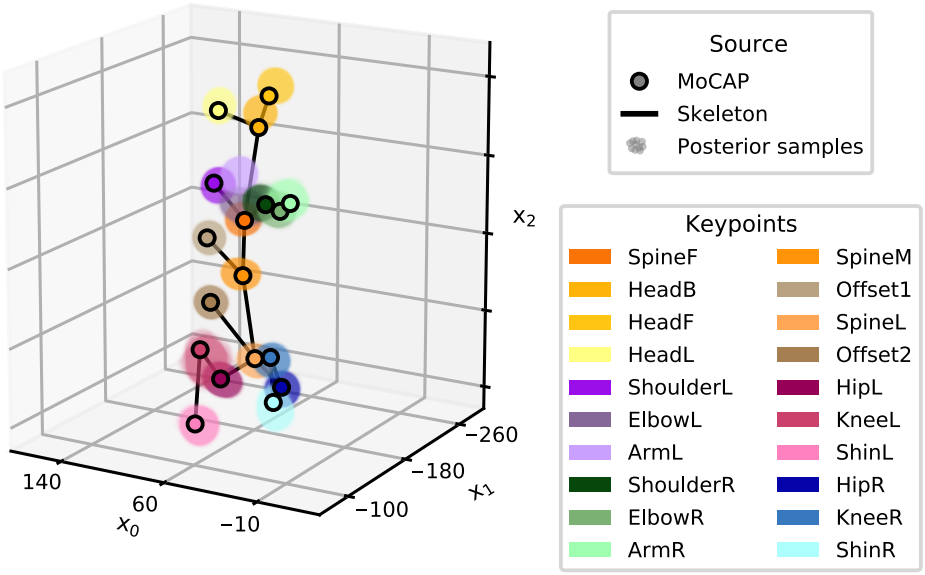

Figure S2: Legend of keypoint marker colors (used in both 2D and 3D plots).

Then, summing over all joints,

$$\begin{aligned}\sum_k \left(\mathrm{R}(h)\,\mu_k\right)^{\mathsf{T}} u_k &= \sum_k \sin(\sphericalangle_{\mathrm{xz}}(\mu_k))\,\sin(\sphericalangle_{\mathrm{xz}}(u_k))\,\cos\left(h - (\sphericalangle_{\mathrm{xy}}(u_k) - \sphericalangle_{\mathrm{xy}}(\mu_k))\right) + \text{const} \\ &= \cos(h)\sum_j \sin(\sphericalangle_{\mathrm{xz}}(\mu_k))\,\sin(\sphericalangle_{\mathrm{xz}}(u_k))\,\cos(\sphericalangle_{\mathrm{xy}}(u_k) - \sphericalangle_{\mathrm{xy}}(\mu_k)) \\ &\quad + \sin(h)\sum_j \sin(\sphericalangle_{\mathrm{xz}}(\mu_k))\,\sin(\sphericalangle_{\mathrm{xz}}(u_k))\,\sin(\sphericalangle_{\mathrm{xy}}(u_k) - \sphericalangle_{\mathrm{xy}}(\mu_k)) + \text{const} \\ &= \cos(h)\,\tau\,\cos(\theta) + \sin(h)\,\tau\,\sin(\theta) + \text{const} \\ &= \tau\cos(h-\theta) + \text{const}.\end{aligned}$$

We have the equalities

$$\begin{aligned}\tau\,\sin(\theta) &= \sum_j \sin(\sphericalangle_{\mathrm{xz}}(u_{t,k}))\,\sin(\sphericalangle_{\mathrm{xz}}(\mu_k))\,\sin(\sphericalangle_{\mathrm{xy}}(u_k) - \sphericalangle_{\mathrm{xy}}(\mu_k)), \\ \tau\,\cos(\theta) &= \sum_j \sin(\sphericalangle_{\mathrm{xz}}(u_{t,k}))\,\sin(\sphericalangle_{\mathrm{xz}}(\mu_k))\,\cos(\sphericalangle_{\mathrm{xy}}(u_k) - \sphericalangle_{\mathrm{xy}}(\mu_k)),\end{aligned}$$

which we use to solve for θ and τ, as in eq. (14).

S4 Implementation details

S4.1 Dataset

We collected 6 hours of data onat 30 Hz from 6 color video cameras and 12 motion capture cameras. The subject was affixed with 20 retroflective markers to collect ground truth 3D data via the motion capture system. 30 minutes of data were withheld for evaluation. The same 192000 unique images were used for all three methods. 2D targets for DeepLabCut were generated by projecting motion capture coordinates into the images as targets.

S4.2 DANNCE

Volumetric representations were constructed from individual images using projective geometry (eq. (1)). Voxels are represented by the RGB values of all pixels whose rays trace to that location. Volums were concatenated

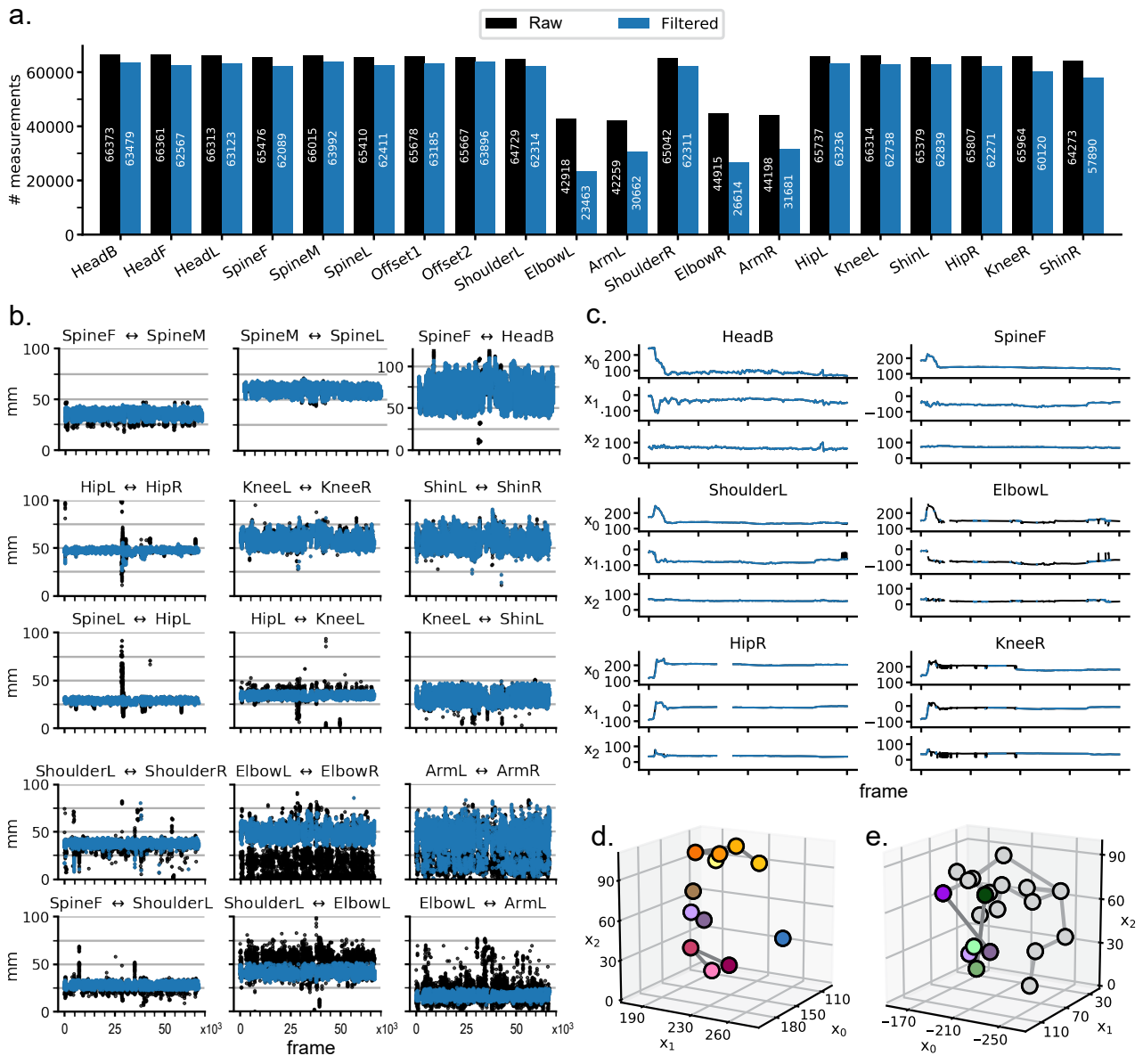

Figure S3: Distances between keypoints used to filter out outlier measurements in MOCAP data. a. Number of frames with a valid measurement by keypoint, in the raw dataset (black, left) and after filtering (blue, right). b. Selected plots of distances between keypoints that are known to be roughly fixed, with distances calculated from the filtered dataset (blue) plotted on top of the distances from the raw dataset (black). Visible black points therefore indicate outlier distances and that one or both of the keypoint measurements were removed in the filtered dataset. c. Selected traces of keypoints from the resulting filtered dataset (blue), plotted on top of the raw measurements (black). d. Example of a frame removed because not enough adjacent keypoints were detected (connected by light grey lines). Here, 10 keypoints are adjacent to at least one other keypoint, out of 12 detected keypoints and 20 total keypoints. Keypoints colored as in Figure S2. e. Example of a frame where left elbow marker (medium purple) was removed because distance between elbow markers (medium purple and medium green) is too close and caused by outlier measurement of left elbow marker. Forelimb keypoints colored as in Figure S2; non-forelimb markers are colored in grey for clarity.

along the color axis for K total views and then inputted into a 3D U-net. The network was implemented in Tensorflow and Keras, and trained using the Adam optimizer. No specific hyperparameter search was performed over the architecture, but deeper networks were found to have higher performance. This method is described in greater detail by Dunn et al. (2020).

S4.3 GIMBAL

We use a subset of the MOCAP ground truth training data to train our model. We choose to filter the training data after observing some consistently poor measurements that are not uncommon in this type of measurement system. For example, when only a subset of keypoints are detected, as in Figure S3d, the measurement system is susceptible to misassignment. Elbow* and Arm* markers frequently fail to be detected (fig. S3a), likely due to self-occlusion and infrared camera positioning. When they are detected, they are prone to being misassigned, as shown in Figure S3b by the large number of frames where raw distances between assigned ElbowL and ElbowR keypoints are much smaller than physically possible. An instance of this is visualized in Figure S3e, where the assigned ElbowL (medium purple) can visually be identified as incorrect.

First, frames with fewer than 50% edges detected are discarded. Then, frames where the distance between elbow keypoints was less than the average distance between shoulder keypoints minus one standard deviation. A similar criteria was applied to the knee keypoints, using the average and standard deviation of the distance between hip keypoints, although Figure S3b shows that this was not necessary for these keypoints. Next, z-scores of the distance between parent and child keypoints were calculated based on the inter-99th percentile of measurements, and measurements that had an absolute z-score greater than 3 were denoted as outliers and removed from the dataset.

Table S1: Prior parameter values used in GIMBAL and its special cases (see main text for description of M0, M1, and M2. Shaded values are used to indicate uninformative priors. `moG` indicates parameters learned from performing expectation maximization (EM) on mixture of Gaussians; `SS` indicates parameters learned from summary statistics; `movMF` indicates parameters learned from performing EM on mixture of vMFs.

		M0	M1	M2	GIMBAL
temporal variance	η_k^2	100	100	50	50
outlier probability	$\beta_{c,k}$	0	`moG`	`moG`	`moG`
outlier variance	$\omega_{1,c,k}^2$	1e6	`moG`	`moG`	`moG`
"inlier" variance	$\omega_{0,c,k}^2$	1e3	`moG`	`moG`	`moG`
root node spatial variance	σ_1^2	1e3	1e3	1e3	1e3
spatial variance	$\sigma_{k>1}^2$	1e3	1e3	`SS`	`SS`
distance	ρ_k	0	0	`SS`	`SS`
direction concentration	$\boldsymbol{\kappa}$	0	0	0	`movMF`
mean direction	$\underline{\boldsymbol{\nu}}$	[1,0,0]	[1,0,0]	[1,0,0]	`movMF`

Robust observation parameters We first calculate the Euclidean error between all 2D observations and the 3D MOCAP measurements projected into their respective plane. Then, we fit a two-mixture Gaussian mixture model to this error data using expectation maximization. The algorithm is initialized with means centered at the origins, inlier covariance $\omega_0^2 = 1$, outlier covariance $\omega_1^2 = 100^2$, and weight $\beta_{1,c,k}$ based on the frequency of errors greater than a roughly picked threshold of 15 px. Figure S4 displays the results of the fitted GMMs for all keypoints for a single camera.

Skeletal parameters After defining a tree-ordering of keypoints, ρ_k is the mean Euclidean distance between a keypoint k and its parent $\pi(k)$, and spatial variance σ_k^2 is the empircal standard deviation of this difference. Summary plots of keypoint distance means and variances from the training data are shown in Figure S5.

Directional priors We perform expectation maximization to fit S mixture of vMFs to observ poses in the the training data. Then at each timestep t, we infer the postural state $s_t \in \{1, \ldots, S\}$, and their corresponding directional means and concentrations for the set of joint, $\{\mu_{s_{t,k}}\}_{k=1}^K, \{\kappa_{s_{t,k}}\}_{k=1}^K$. Conditioned on this state s_t, we assume that the distribution of the joint direction vectors are conditionally independent of each other.

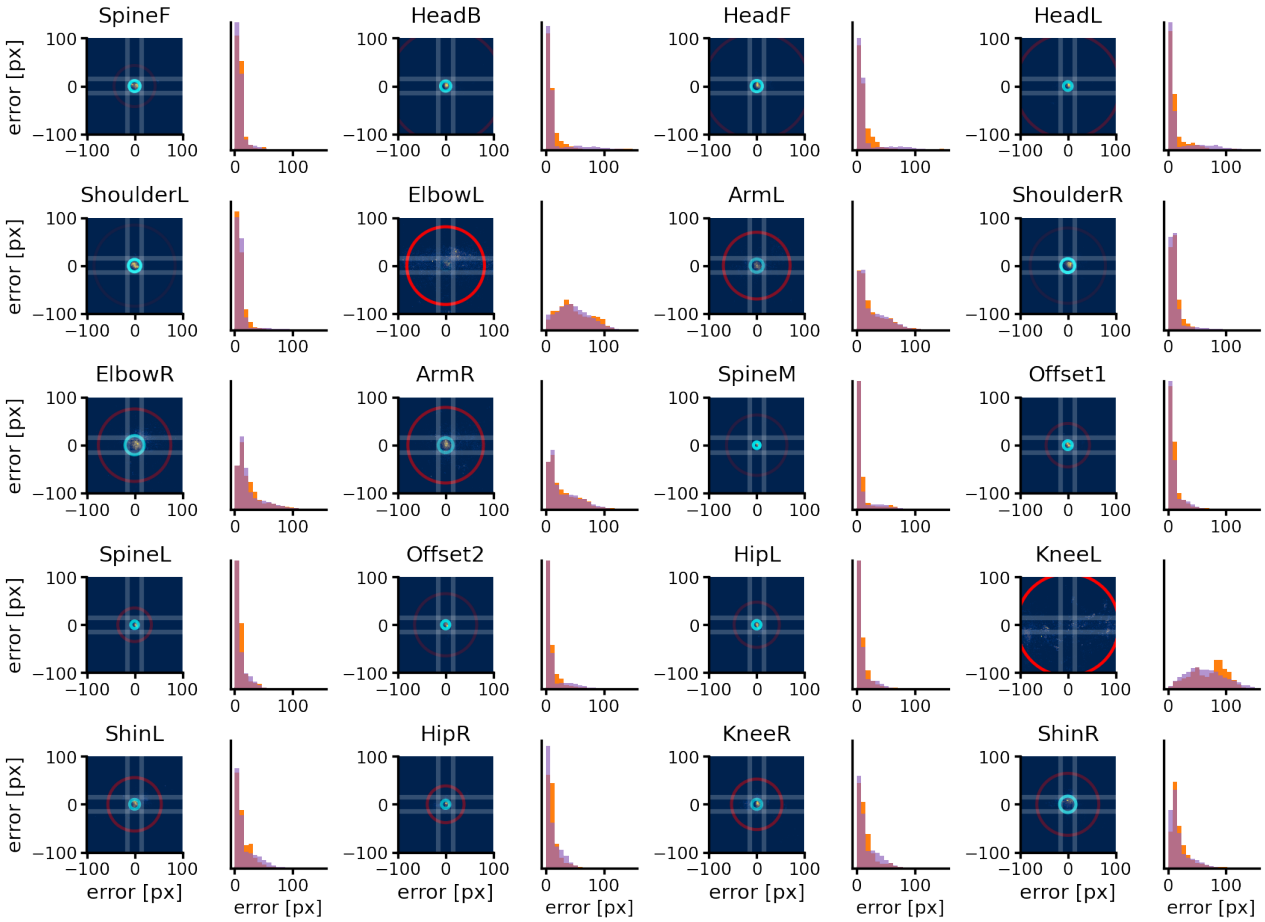

Figure S4: Error distributions of 2D DLC observations from training data for all keypoints from a representative camera. In the color plots on the left, circles denote 2 standard deviation boundary of fitted Gaussian distributions (inlier: cyan, outlier: red). Opacity of line is proportional to the fitted weight of the respective distribution. In the histograms on the right, orange density represents the empirical error distribution. Purple density represents samples from the fitted Gaussian mixture model.

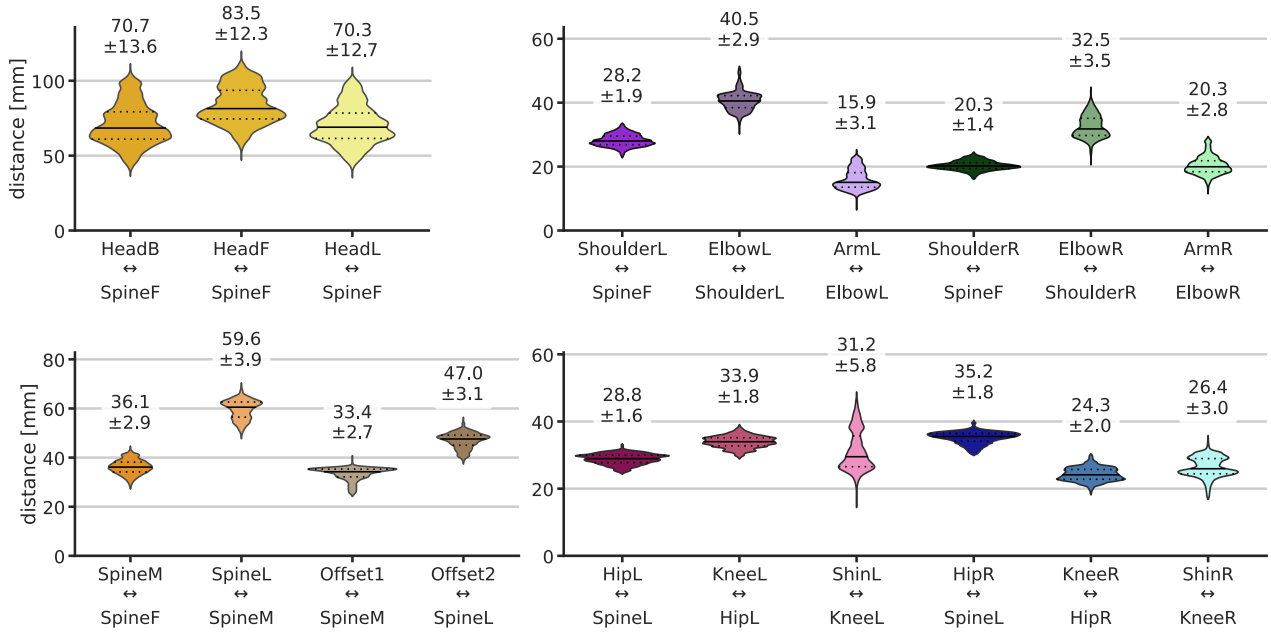

Figure S5: Skeletal parameters summarized from training data. Violin plots of distances between keypoints k and their parent $\pi(k)$, as defined by graph $\mathcal{G}$. The profile of the violin plots indicate that in many cases, the error is not strictly Gaussian, or unimodal. The large variance in the distances between Head* and SpineF markers is due to a violation in our approximate rigidty assumption. Variance can be reduced by inferring an (unobserved) keypoint situated at the base of the head, which would serve as an auxiliary joint between the Head* and SpineF markers. Annotations denote the mean ± standard deviation.

$$s_{t,s} \sim \text{Categorical}(\pi) \qquad \text{for } s = 1, \ldots, S$$
$$u_{t,k} \sim \text{vMF}(\mu_{s_t,k}, \kappa_{s_t,k}) \qquad \text{for } k = 1, \ldots, K$$

Our expectation step involves a product over all the vMF distributions,

$$\mathbb{E}[s_{ts}] = \frac{\pi_s \prod_{k=1}^{K} \text{vMF}(u_{tk} \mid \nu_{s_t k}, \kappa_{s_t k})}{\sum_{k'} \pi_{k'} \text{vMF}(u_{tk} \mid \mu_{k'k}, \kappa_{k'k})} \tag{11}$$

Our maximization step

$$\pi_k = \frac{1}{N} \sum_i \mathbb{E}[s_{ts}] \tag{12}$$

$$\mu_{sk} = \frac{R_{skj}}{\|R_{sk}\|} \qquad \text{where } R_{sk} = \sum_t \mathbb{E}[s_{ts}]\, u_{tk} \tag{13}$$

$$\kappa_{sk} = A_p^{-1}(\bar{r}_{sk}) \qquad \bar{r}_{sk} = \frac{1}{\sum_i \mathbb{E}[s_{ts}]} \|R_{sk}\| \tag{14}$$

We estimate the concentration parameter using an Amos-type bound (Hornik and Grün, 2014).