Supplementary Material

This document accompanies the main manuscript, comprises of supporting derivation and additional results.

1 Derivation of Latent Filter

We assume a generative model with an underlying latent dynamical system with

$$p(\mathbf{o}_{1:T}|\mathbf{a}_{1:T}) = \int p(\mathbf{o}_{1:T}|\mathbf{s}_{1:T}, \mathbf{a}_{1:T})\, p(\mathbf{s}_{1:T}|\mathbf{a}_{1:T})\, d\mathbf{s}_{1:T}$$
$$= \int \prod_{t=1}^{T} p(\mathbf{o}_t|\mathbf{o}_{1:t-1}, \mathbf{s}_t, \mathbf{s}_{1:t-1}, \mathbf{a}_{1:t})\, p(\mathbf{s}_t|\mathbf{o}_{1:t-1}, \mathbf{s}_{t-1}, \mathbf{s}_{1:t-2}, \mathbf{a}_t, \mathbf{a}_{1:t-1})\, d\mathbf{s}_{1:t} \tag{1}$$

Eq. 1 becomes computationally expensive as the number of time steps increases, due to the growing complexity of conditional dependencies. To address this, we adopt the first-order Markov assumption, which states that (i) the latent state at time t depends only on the immediate past state $\mathbf{s}_{t-1}$ (rather than on the full history $\mathbf{s}_{1:t-1}$) and that (ii) the observations at each time step are fully determined by the current state. This assumption enables a more compact factorization of the joint distribution and facilitates tractable learning and inference over long temporal sequences

$$p(\mathbf{o}_t|\cancel{\mathbf{o}_{1:t-1}}, \mathbf{s}_t, \cancel{\mathbf{s}_{1:t-1}}, \cancel{\mathbf{a}_{1:t}}) = p(\mathbf{o}_t|\mathbf{s}_t) \tag{2}$$
$$p(\mathbf{s}_t|\cancel{\mathbf{o}_{1:t-1}}, \mathbf{s}_{t-1}, \cancel{\mathbf{s}_{1:t-2}}, \mathbf{a}_t, \cancel{\mathbf{a}_{1:t-1}}) = p(\mathbf{s}_t|\mathbf{s}_{t-1}, \mathbf{a}_t)$$

resulting in the simplified generative model:

$$p(\mathbf{o}_{1:T}|\mathbf{a}_{1:T}) = p(\mathbf{o}_1|\mathbf{s}_1, \mathbf{a}_1)\, p(\mathbf{s}_1) \int \prod_{t=2}^{T} p(\mathbf{o}_t|\mathbf{s}_t)\, p(\mathbf{s}_t|\mathbf{s}_{t-1}, \mathbf{a}_t)\, d\mathbf{s}_t\,. \tag{3}$$

With the proposed factorization results in the following generative model and factorization from Eq.3

$$p(\mathbf{o}_{1:T}|\mathbf{a}_{1:T}) = \iint \prod_{t=1}^{T} p(\mathbf{o}_t|\mathbf{z}_t)\, p(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}_t, \mathbf{a}_t)\, p(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{a}_{t-1})\, d\mathbf{y}_{t-1} d\mathbf{z}_{t-1} \tag{4}$$

To compute the observation likelihood, we introduce variational distribution $q_\theta(\mathbf{z}_{1:T}, \mathbf{y}_{1:T}) \sim p(\mathbf{z}_{1:T}, \mathbf{y}_{1:T}|\mathbf{o}_{1:T}, \mathbf{a}_{1:T})$. The *Evidence Lower Bound Objective* (ELBO) for the generative model in Eq. 4 is formulated from the KL Divergence inequality:

$$D_{\mathrm{KL}} = -\iint q_\theta(\mathbf{z}_{1:T}, \mathbf{y}_{1:T}|\mathbf{o}_{1:T}, \mathbf{a}_{1:T}) \log \left[\frac{p(\mathbf{z}_{1:T}, \mathbf{y}_{1:T}|\mathbf{o}_{1:T}, \mathbf{a}_{1:T}}{q_\theta(\mathbf{z}_{1:T}, \mathbf{y}_{1:T}|\mathbf{o}_{1:T}, \mathbf{a}_{1:T}}\right] d\mathbf{z}_{1:T}\, d\mathbf{y}_{1:T} \geq 0 \tag{5}$$

resulting in the following objective function

$$\begin{aligned} \log p(\mathbf{o}_{1:T}|\mathbf{a}_{1:T}) \geq\ & \mathbb{E}_{q_\theta(.)}[\log p(\mathbf{o}_{1:T}|\mathbf{z}_{1:T}, \mathbf{y}_{1:T}, \mathbf{a}_{1:T})] \\ & -\mathbb{E}_{q_\theta(.)}\left[\log\left(\frac{q_\theta(\mathbf{z}_{1:T}, \mathbf{y}_{1:T}|\mathbf{o}_{1:T}, \mathbf{a}_{1:T})}{p(\mathbf{z}_{1:T}, \mathbf{y}_{1:T}|\mathbf{a}_{1:T})}\right)\right] . \end{aligned} \tag{6}$$

Re-introducing Markov assumption and simplification presented in the generative model in the inference into the regularization term of the ELBO

$$q_\theta(\mathbf{z}_{1:T}, \mathbf{y}_{1:T}|\mathbf{o}_{1:T}, \mathbf{a}_{1:T}) = q_\theta(\mathbf{z}_1|\mathbf{y}_1)\, q_\psi(\mathbf{y}_1) \prod_{t=2}^{T} q_\theta(\mathbf{z}_t, \mathbf{y}_t|\mathbf{z}_{t-1}, \mathbf{y}_{t-1}, \mathbf{o}_{1:t}, \mathbf{a}_{1:t}) \tag{7}$$

$$\begin{aligned} q_\theta(\mathbf{z}_t, \mathbf{y}_t|\mathbf{z}_{t-1}, \mathbf{y}_{t-1}, \mathbf{o}_{1:t}, \mathbf{a}_{1:t}) &= \frac{p(\mathbf{o}_t|\mathbf{z}_t)\, q_\theta(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}_t, \mathbf{a}_t) q_\psi(\mathbf{y}_t|\mathbf{z}_{t-1}, \mathbf{y}_{t-1}, \mathbf{a}_{t-1})}{\iint p(\mathbf{o}_t|\mathbf{z}_t)\, p(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}_t, \mathbf{a}_t)\, d\mathbf{z}_{t-1} d\mathbf{y}_t} \\ &\sim \eta\, \underbrace{q_\psi(\mathbf{z}_t|\mathbf{o}_t)\, q_\theta(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}_t, \mathbf{a}_t)}_{q^{filt}(\mathbf{z}_t)} q_\psi(\mathbf{y}_t|\mathbf{z}_{t-1}, \mathbf{y}_{t-1}, \mathbf{a}_{t-1}) \end{aligned} \tag{8}$$

where the denominator is the normalizer, results in the simplified ELBO:

$$\begin{aligned} \mathcal{F}_{\mathrm{ELBO}}(\theta, \psi) = \mathbb{E}_{q_\psi(.)}\left[\sum_{t=1}^{T} \log p(\mathbf{o}_t|\mathbf{z}_t)\right] &- \sum_{t=2}^{T} \mathrm{KL}[q^{filt}(\mathbf{z}_t)||p(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}_t, \mathbf{a}_t) \\ &- \sum_{t=2}^{T} \mathrm{KL}[q_\psi(\mathbf{y}_t|.)||p(\mathbf{y}_t|\mathbf{a}_t)] \end{aligned} \tag{9}$$

2 Additional Results

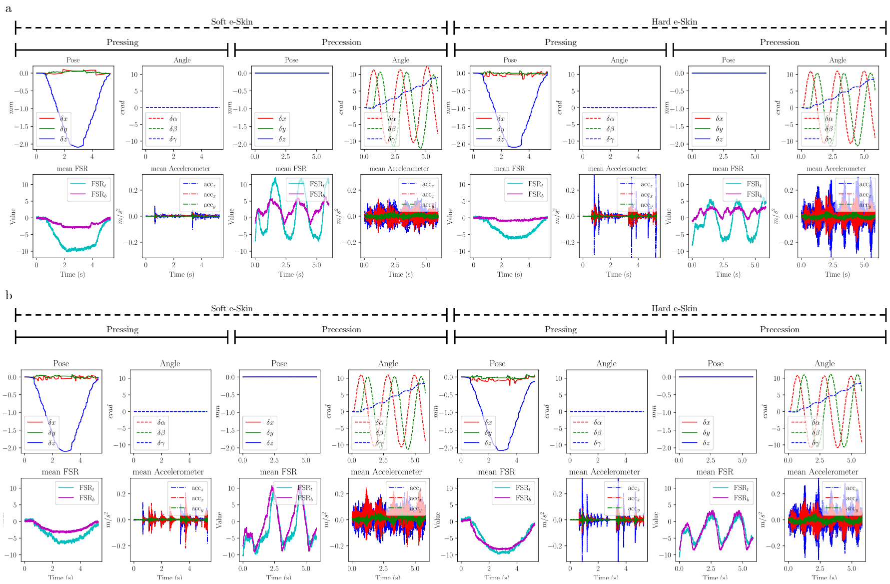
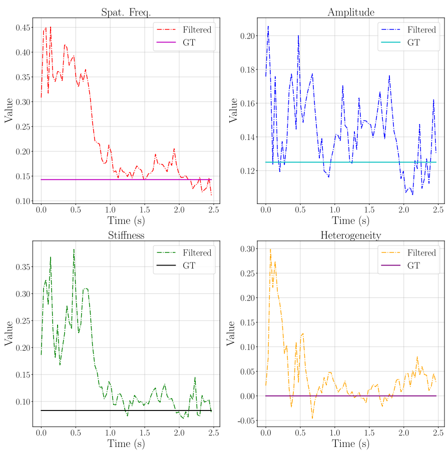
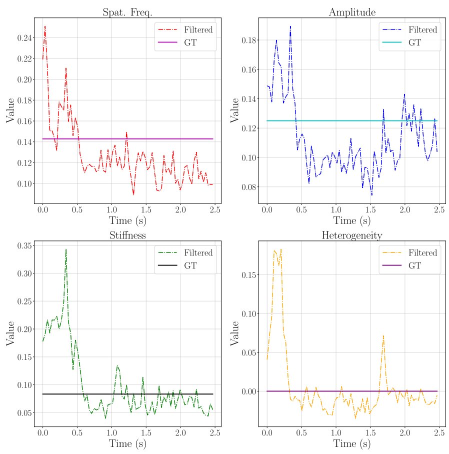


Fig. 1: Illustrative samples of the raw tactile observations obtained during the experiments. a) is Object Id 0, while b) is Object Id 26.



(a) $A_{\alpha,\beta} = 6\,mm, \omega_{\alpha,\beta} = 0.8\,Hz$.



(b) $A_{\alpha,\beta} = 3\,mm, \omega_{\alpha,\beta} = 0.2\,Hz$.

Fig. 2: Illustrative example of latent filtering for Object ID O, under the different action parameters for precession interaction.

Method	E-Skin	Clust.		Class.	Trajectory NMSE				
		S.Score	DB	LogReg	S. Freq.	Ampl.	Stiff.	Hetero.	Overall
Base. I	Soft	-0.15	5.73	0.77	NA	NA	NA	NA	NA
Base. I	Hard	-0.13	5.08	0.67	NA	NA	NA	NA	NA
Base. II	Soft	-0.11	20.65	0.05	0.44	0.20	0.46	0.32	0.36
Base. II	Hard	-0.36	36.71	0.08	0.46	0.20	0.47	0.27	0.35
Base. III	Soft	0.04	3.17	0.83	0.11	0.05	0.13	0.10	0.10
Base. III	Hard	0.00	3.32	0.79	0.12	0.05	0.13	0.12	0.10
MULTI-L	Soft	0.46	0.82	0.99	0.06	0.02	0.08	0.09	0.06
MULTI-L	Hard	0.45	0.84	0.98	0.04	0.02	0.07	0.10	0.06
Method	**Inter.**	**Clust.**		**Class.**	**Trajectory NMSE**				
		S.Score	DB	LogReg	S. Freq.	Ampl.	Stiff.	Hetero.	Overall
ACC	Pressing	-0.20	11.88	0.15	0.14	0.06	0.23	0.26	0.17
ACC	Precision	-0.08	7.61	0.15	0.14	0.06	0.22	0.27	0.17
ACC	Slipping	-0.12	14.31	0.16	0.16	0.07	0.24	0.28	0.18
FSR_t	Pressing	0.32	1.09	0.97	0.07	0.03	0.11	0.13	0.08
FSR_t	Precision	0.20	1.58	0.91	0.08	0.03	0.13	0.16	0.10
FSR_t	Slipping	-0.05	3.75	0.73	0.13	0.04	0.18	0.24	0.15
FSR_b	Pressing	0.30	1.17	0.93	0.07	0.02	0.11	0.14	0.09
FSR_b	Precision	0.14	1.93	0.87	0.07	0.02	0.11	0.16	0.09
FSR_b	Slipping	-0.09	4.60	0.61	0.11	0.04	0.17	0.26	0.14
FSR_{tb}	Pressing	0.39	0.96	0.97	0.05	0.02	0.08	0.12	0.07
FSR_{tb}	Precision	0.22	1.55	0.96	0.05	0.02	0.10	0.13	0.07
FSR_{tb}	Slipping	-0.05	2.78	0.78	0.11	0.04	0.17	0.22	0.13
MULTI-E	Pressing	0.36	1.01	0.97	0.06	0.02	0.09	0.13	0.07
MULTI-E	Precision	0.21	1.64	0.96	0.06	0.02	0.09	0.14	0.08
MULTI-E	Slipping	0.14	1.62	0.95	0.06	0.02	0.09	0.14	0.08
MULTI-L	Pressing	0.46	0.82	0.99	0.06	0.02	0.08	0.09	0.06
MULTI-L	Precision	0.33	1.24	0.99	0.06	0.02	0.08	0.08	0.06
MULTI-L	Slipping	0.23	1.40	0.98	0.07	0.03	0.11	0.13	0.08

Fig. 3: Quantitative evaluation results. S.Score refers to the Silhouette L2 score (range: -1 to 1, higher is better) and DB denotes the Davies–Bouldin score (lower is better); both are clustering metrics. LogReg indicates classification performance (range: 0–1, higher is better). S.Freq, Ampl., Stiff., and Hetero correspond to the mechanical properties of the objects: spatial frequency, amplitude, stiffness, and heterogeneity, respectively. Trajectory NMSE represents the normalized mean squared error of predicted trajectories (lower is better). The table facilitates visual aide, with brighter color signifying worst performance across different comparative methods.

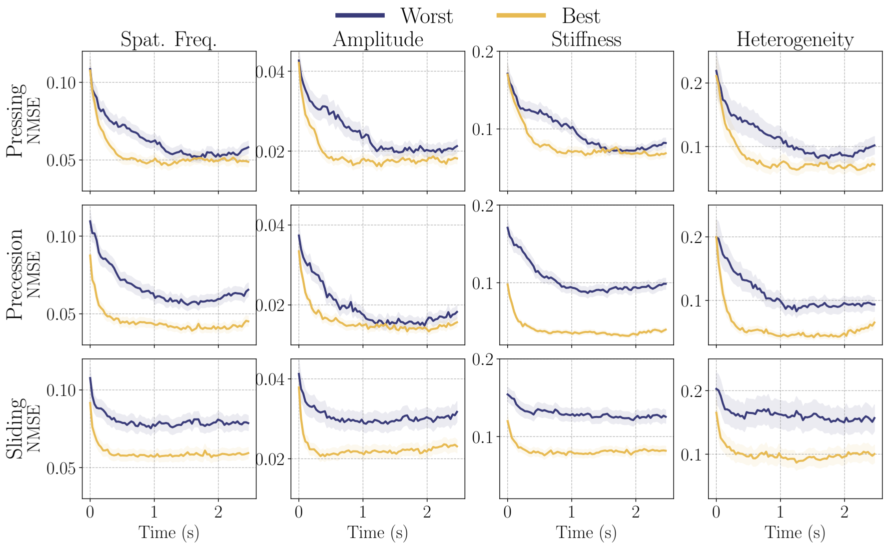


Fig. 4: Comparison of different action parameters for MULTI-L (Soft) approach. The best and worst performing interaction parameters are selected to avoid visual and interpretive clutter.

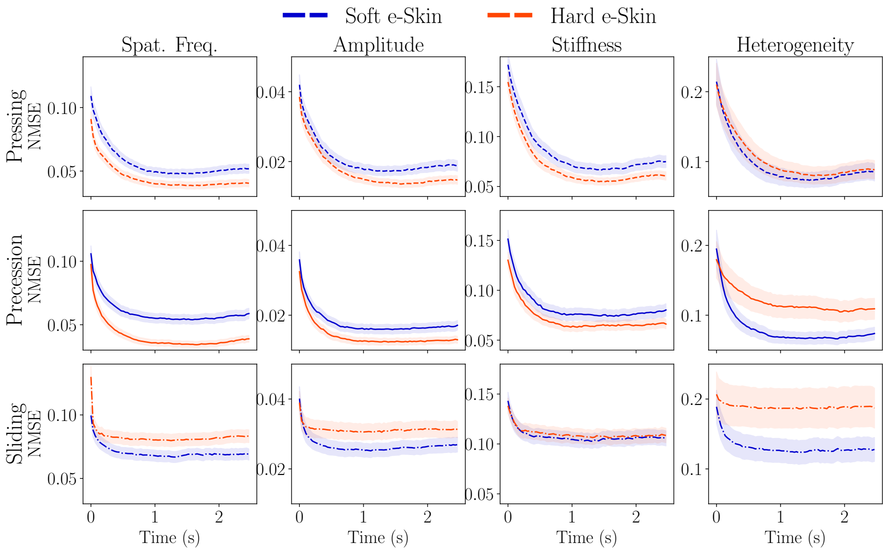


Fig. 5: Comparison between hard and soft skin types using MULTI-L (multi-modal late fusion) approach.