Appendix A Implementation Details

A.1 TactileVAD Implementation and Training Details

Algorithm 1: TactileVAD Training Algorithm

Input:
K: Number of samples
T: Number of timesteps
$(\mathbf{u}_0, \mathbf{u}_1, \dots \mathbf{u}_{T-1})$: Initial control sequence

$f_{\text{lat\_dyn}} : \{\mathbf{A}^{(0)}, \mathbf{B}^{(0)}\} \leftarrow$ Initialization;
$f_{\text{dec}} : \{\theta_i^{(0)}\} \leftarrow$ Initialization;
$q : \{\boldsymbol{\mu}_i^{(0)}, \boldsymbol{\Sigma}_i^{(0)}\} \leftarrow$ Initialization;
$k \leftarrow 0$;
while *not converged* **do**
  $[\mathbf{z}_t] \sim q(\mathbf{z}_t; \boldsymbol{\mu}_t^{(k)}, \boldsymbol{\Sigma}_t^{(k)})$;
  $q(\hat{\mathbf{z}}_{t+1}; \hat{\boldsymbol{\mu}}_t^{(k)}, \hat{\boldsymbol{\Sigma}}_t^{(k)}) = \mathcal{N}(\mathbf{A}\boldsymbol{\mu}_t + \mathbf{B}\mathbf{u}_t, \mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}^\top)$;
  $[\hat{\mathbf{z}}_{t+1}] \sim q(\hat{\mathbf{z}}_{t+1}; \hat{\boldsymbol{\mu}}_{t+1}^{(k)}, \hat{\boldsymbol{\Sigma}}_{t+1}^{(k)})$;
  $[\mathbf{x}_t] = [f_{\text{dec}}(\mathbf{z}_{t+1})]$;
  $[\hat{\mathbf{x}}_{t+1}] = [f_{\text{dec}}(\hat{\mathbf{z}}_{t+1})]$;
  $l_{\text{rec}} = \mathbb{E}_{\mathbf{z}_t \sim q_t}[-\log P_\theta(\mathbf{x}_t|\mathbf{z}_t)]$;
  $l_{\text{[pred]}} = \mathbb{E}_{\hat{\mathbf{z}}_{t+1} \sim \hat{q}_{t+1}}[-\log P_\theta(\mathbf{x}_{t+1}|\hat{\mathbf{z}}_{t+1})]$;
  $l_{\text{cons}} = \text{KL}(q(\hat{\mathbf{z}}_{t+1}; \hat{\boldsymbol{\mu}}_t^{(k)}, \hat{\boldsymbol{\Sigma}}_t^{(k)}) \| q(\mathbf{z}_t; \boldsymbol{\mu}_{t+1}^{(k)}, \boldsymbol{\Sigma}_{t+1}^{(k)}))$;
  $l_{\text{kl}} = \text{KL}(q(\mathbf{z}_t; \boldsymbol{\mu}_{t+1}^{(k)}, \boldsymbol{\Sigma}_{t+1}^{(k)}) \| P(Z))$;
  $l = \alpha_{\text{rec}} l_{\text{rec}} + \alpha_{\text{pred}} l_{\text{pred}} + \alpha_{\text{cons}} l_{\text{cons}} + \alpha_{\text{kl}} l_{\text{kl}}$;
  $\{\theta, \mathbf{A}, \mathbf{B}, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1}\}^{(k+1)} \leftarrow \underset{\theta, \mathbf{A}, \mathbf{B}, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1}}{\text{grad\_step}} (l)$;
  $k \leftarrow k + 1$;

A.2 TactileVAD Control Details

Algorithm 2: TactileVAD Control Algorithm

Input:
$\mathbf{x}_g$: Goal observation representing $\mathbf{s}_g$
$\mathbf{Q}$: quadratic state cost
$\mathbf{R}$: quadratic action cost
n: number of inference iterations

$\mathbf{K} \leftarrow \text{LQRSolution}(\mathbf{Q}, \mathbf{R})$;
$\mathbf{z}_g \leftarrow \text{inference}_{f_{\text{dec}}}(\mathbf{x}_g)$;
while *not converged* **do**
  $\mathbf{x}_t \leftarrow \text{get\_observation}()$;
  $\mathbf{z}_t \leftarrow \text{inference}_{f_{\text{dec}}}(\mathbf{x}_t)$;
  $\mathbf{u}_t \leftarrow -\mathbf{K}(\mathbf{z}_t - \mathbf{z}_g)$;
  $\text{send\_to\_actuator}(\mathbf{u}_t)$;
  $k \leftarrow k + 1$;

Appendix B Experimental Details

B.1 Baselines Details

All four benchmarked models share the same decoder architectures and latent dynamics parametrization. Encoder-decoder approaches (AE and E2C) share the encoder structure but E2C extends it to produce both

Method	Bubbles Final Imprint Errors ($\cdot 10^{-6}$)						Gelslim Final Imprint Errors					
	20mm Rod (Train)		15mm Rod		30mm Rod		20mm Rod (Train)		15mm Rod		30mm Rod	
	Mean ↓	Std ↓	Mean ↓	Std ↓	Mean↓	Std↓	Mean ↓	Std ↓	Mean ↓	Std ↓	Mean↓	Std↓
AE	8.6062	4.3312	2.3818	2.1589	15.269	12.121	30.009	9.90	27.22	4.17	18.84	8.319
E2C	5.9678	5.2829	4.4046	4.6345	3.9464	2.0997	25.28	7.576	29.20	7.27	19.05	5.79
AD	6.5889	3.4163	3.3810	1.9211	11.281	3.8071	21.076	3.75	29.24	11.45	15.98	1.51
VAD (ours)	**2.3723**	**1.0926**	**0.6295**	**0.2842**	**3.1057**	**2.67**	**19.631**	**12.252**	**23.91**	**9.44**	**14.52**	**6.34**

Table 4: **Tactile Rod Grasping Evaluation: (Tactile similarity)**

$\boldsymbol{\mu}$ and diag($\boldsymbol{\Sigma}$). Fig. B.1 shows the schematic for E2C over trajectories. Fig. B.2 shows the auto-decoder architecture.

B.2 Moving Block Control Details

The moving block task is defined by 20×20 grid that contains a 3×12 block. The sensed area is 12×8 and is centered in the middle of the space. The true block state is defined by the block top coorner coordinates (x, y) The block motion is defined by limited on a box of size $u = (\delta x, \delta y) \in \mathcal{U} \subseteq [-3, 3] \times [-3, 3]$. States are encoded as binary masks The pose error is computed as the Manhattan distance. The models were trained on 500 random trajectories of length 20 steps. The reconstruction and prediction loss over observations is computed as the BCE against the ground truth states.

B.3 Tactile Rod Grasping Details

The rod grasping task is defined over a $SE(2)$ space of robot-rod configurations. The space limits are $\mathcal{S} \subseteq [-20, 20] \times [-20, 50] \times [-\frac{\pi}{2}, \frac{\pi}{2}]$. Robot action are constrained within a box defined as:

$$\mathbf{u} = (\delta x, \delta y, \delta\theta) \in \mathcal{U} \subseteq [-6, 6] \times [-5, 5] \times [-0.09\pi, 0.09\pi]$$

For the bubbles sensors, states are encoded as deformation depthmaps of size $(2, 25, 20)$. Gelslim data is encoded as color differences encoded as grayscale of size $(2, 20, 20)$

$$\text{Pose Score}(\begin{bmatrix} x_1 \\ y_1 \\ \theta_1 \end{bmatrix}, \begin{bmatrix} x_2 \\ y_2 \\ \theta_2 \end{bmatrix}) = \text{MSE}(x_1, x_2) + \text{MSE}(y_1, y_2) + r_g \text{MSE}(\theta_1, \theta_2) \tag{5}$$

where r_g is the radius of gyration of the rod.

Appendix C Additional Experiment Results

C.1 Tactile Rod Grasping

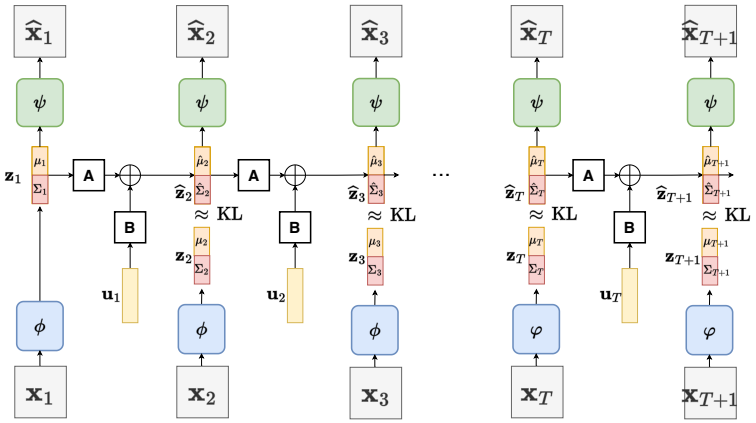

Figure B.1: **E2C Model**

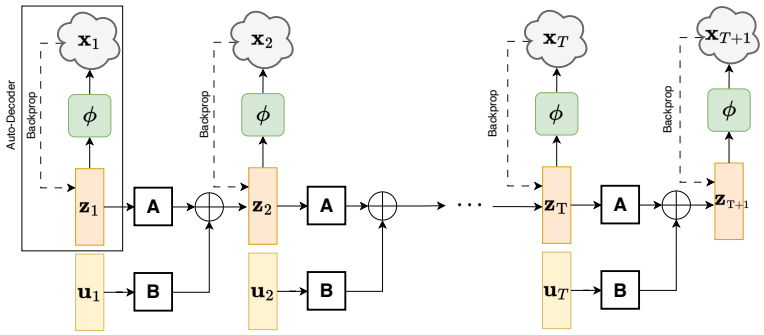

Figure B.2: **Auto-Decoder Latent Linear Dynamics Model (AD)**