APPENDIX A
DETAILS ON STATE SPACE FORMULATION

This section provides a detailed formulation of the system's state space, encompassing the key variables that govern the robotic in-hand writing process:

- **Robotic Joint** ($\mathbf{x}_{\text{wt}} \in \mathbb{R}^3$, $\mathbf{x}_{\text{fg}} \in \mathbb{R}^6$): Mimicking human writing, the wrist position $\mathbf{x}_{\text{wt}}$ remains largely stationary, adjusting mainly for height, with left-right translations occurring only when transitioning between words. Most movements are managed by the fingers, with their joints normalized to [0, 1] as $\mathbf{x}_{\text{fg}}$, based on their respective lower and upper limits.
- **In-hand Contact** ($\mathbf{x}_{\text{int}} \in \mathbb{R}^{3\times3}$, $\mathbf{f}_{\text{int}} \in \mathbb{R}^3$): For each of the three fingers, in-hand contact positions $\mathbf{x}^i_{\text{int}}(i = 1, 2, 3)$ are obtained by averaging the positions of tactile taxels detecting pressure above a predefined threshold. The corresponding in-hand contact forces $\mathbf{f}_{\text{int}}$ are calculated by vector summing forces from these active sensors. Both $\mathbf{x}_{\text{int}}$ and $\mathbf{f}_{\text{int}}$ are first expressed in the local coordinate frame of each finger link and subsequently transformed into the global coordinate system.
- **Writing Tool Pose** ($\mathbf{p}_{\text{obj}} \in \mathbb{R}^3$): The writing tool's pose is represented by a unit vector $\mathbf{p}_{\text{obj}}$, with the orientation around its major axis neglected. The initial object pose is denoted as $\mathbf{p}_{\text{prior}}$.
- **Target Trajectory** ($\mathbf{x}_{\text{tg}} \in \mathbb{R}^2$, $\mathbf{v}_{\text{tg}} \in \mathbb{R}^2$): The system follows a desired trajectory of target positions $\mathbf{x}_{\text{tg}}$ on the writing plane, with corresponding velocities $\mathbf{v}_{\text{tg}}$ tracked to ensure smooth motion.
- **Extrinsic Contact** ($\mathbf{x}_{\text{ext}} \in \mathbb{R}^3$, $\mathbf{f}_{\text{ext}} \in \mathbb{R}$): Extrinsic contact positions $\mathbf{x}_{\text{ext}}$ are estimated using in-hand positions $\mathbf{x}_{\text{int}}$ and the height of writing surface. While tangential friction is treated as a disturbance, only the normal component of the extrinsic contact force $\mathbf{f}_{\text{ext}}$ is considered.

These variables collectively define the system's state space, which is fundamental for modeling both in-hand and extrinsic contact dynamics in the robotic writing process.

APPENDIX B
DETAILS ON REINFORCEMENT LEARNING POLICY

This section provides a comprehensive overview of the reinforcement learning (RL) policy used for finger motion control in the robotic in-hand writing system.

The finger motion control problem is modeled as a finite-horizon discounted Markov decision process (MDP), where a robotic hand, acting as an agent, interacts with a stochastic environment by selecting sequential actions. A deep reinforcement learning (RL) algorithm optimizes the control policy to maximize task performance. The MDP consists of a state space ($\mathcal{S}$), representing all possible system states, and an action space ($\mathcal{A}$), representing all possible control actions. The state transition function $T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ defines the probability distribution of state transitions given an action. The reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ assigns a scalar reward for each state-action pair. At each time step t, the policy π observes the current state $s_t \in \mathcal{S}$, selects an action $a_t = \pi(s_t) \in \mathcal{A}$, and receives a reward $r(s_t, a_t)$. The objective is to maximize the expected cumulative discounted reward:

$$\mathbb{E}\pi \left[\sum t = 0^T \gamma^t r(s_t, a_t) \right], \tag{A1}$$

where $\gamma \in [0, 1)$ is the discount factor that balances immediate and future rewards.

1) Observation Space: We choose Proximal Policy Optimization (PPO) due to its stability and parallelization capabilities, making it ideal for real-time manipulation tasks. The RL agent operates within a carefully designed observation space that captures critical task-related features, providing essential information for finger motion control. The observation space includes the object vector, extrinsic contact forces and positions, the target's relative position and velocity, normalized finger joint positions, and in-hand contact forces and positions at the fingertips, as outlined in TABLE A1.

TABLE A1
OBSERVATION SPACE

Type	State Variable
Observable	
Finger Joint Positions	$\mathbf{x}_{\text{fg}} \in \mathbb{R}^6$
Target Position and Velocity	$\mathbf{x}_{\text{tg}} \in \mathbb{R}^2$, $\mathbf{v}_{\text{tg}} \in \mathbb{R}^2$
In-hand Contact Force and Position	$\mathbf{x}_{\text{int}} \in \mathbb{R}^{3\times3}$, $\mathbf{f}_{\text{int}} \in \mathbb{R}^3$
Unobservable	
Object Vector	$\mathbf{p}_{\text{obj}} \in \mathbb{R}^3$
Extrinsic Contact Force and Position	$\mathbf{x}_{\text{ext}} \in \mathbb{R}^3$, $\mathbf{f}_{\text{ext}} \in \mathbb{R}$

2) Action Space: The action space corresponds to the desired displacements of six finger joints. A Proportional-Derivative (PD) control scheme computes target joint positions by blending previous positions with the action input, scaled by the maximum allowable joint velocity and simulation timestep. To ensure stability, the target position is clipped within joint limits, and the low-level PD controller is applied with proportional and derivative gains to minimize position tracking error.

3) Reward Design: The reward function consists of multiple components, each designed to guide the agent toward stable and effective manipulation. These reward components encourage precise task execution, smooth actions, and consistent contact with the object and external surfaces. The reward components are summarized in TABLE A2.

TABLE A2
REWARD COMPONENTS AND THEIR FORMULAS

Reward	Formula
r_{pos}	$-\log_{10}(\lVert\mathbf{x}_{\text{ext},xy} - \mathbf{x}_{\text{tg}}\rVert{+}3)$
r_{height}	$-\log_{10}(\lVert\mathbf{x}_{\text{ext},z} - z_{\text{plane}}\rVert{+}3)$
r_{time}	0.5
r_{smooth}	$-0.01 \cdot \lVert\mathbf{a}_{\text{t}} - \mathbf{a}_{\text{t-1}}\rVert$
r_{con}	$\begin{cases} 0.001 \cdot \sum \mathbf{f}_{\text{int}}, & \text{if all contact exists} \\ -0.05, & \text{if any fingertip has no contact} \\ -0.5, & \text{if all contact points are zero} \end{cases}$
r_{ext}	$0.08 \cdot (-\lVert\mathbf{f}_{\text{ext}} - f_{\text{th}}\rVert)$

Appendix C
Details on Sim-to-real transfer

This section provides an in-depth description of the sim-to-real transfer processes, including tactile sensor calibration, Covariance Matrix Adaptation Evolution Strategy (CMA-ES) optimization for finger joints, and domain randomization parameterization.

1) Tactile Sensor Calibration: Accurate contact force measurements are essential for simulating realistic tactile feedback. The tactile sensor calibration process maps the voltage output v_{tac} from each taxel to the corresponding applied force F_{tac}. The calibration platform consists of a SCARA-type three-axis robotic arm and a three-axis gimbal. At the end of the arm, an ATI mini45 force/torque sensor and a cylindrical pressing structure are used to apply controlled pressure to each taxel, as shown in Fig. A1 (a). Normal forces ranging from 0 to 2.5 N are applied, with three repeated measurements taken for each taxel. To establish force mapping with a dead zone, we fit a nonlinear function to each taxel's response using the Levenberg-Marquardt (LM) algorithm. The relationship between the piezoelectric output v_{tax} and corresponding normal force F_{tac} for the i-th taxel is modeled as:

$$F_{\tilde{\text{tac}}} = a \cdot \frac{b v_{\tilde{\text{tax}}}}{1 - b v_{\tilde{\text{tax}}}} + c, \tag{A2}$$

where $F_{\tilde{\text{tac}}} = F_{\text{tac}}/2.5$ represents the normalized measured force, $v_{\tilde{\text{tax}}} = v_{\text{tax}}/18000$ is the sensor's normalized output value. Outliers caused by off-center presses or irregularities are filtered to improve consistency. As shown in Fig. A1 (b), calibrating taxel 71 with a dead zone yields parameters $a = 0.1$, $b = 1.08$, and $c = 0.16$, with an average error of 1.63%. Without the dead zone, the parameters are $a = 0.37$, $b = 0.82$, and the error increases to 7.16%. The inclusion of a dead zone largely improves accuracy, with most taxels exhibiting an error below 5%, leading to more reliable contact force measurements.

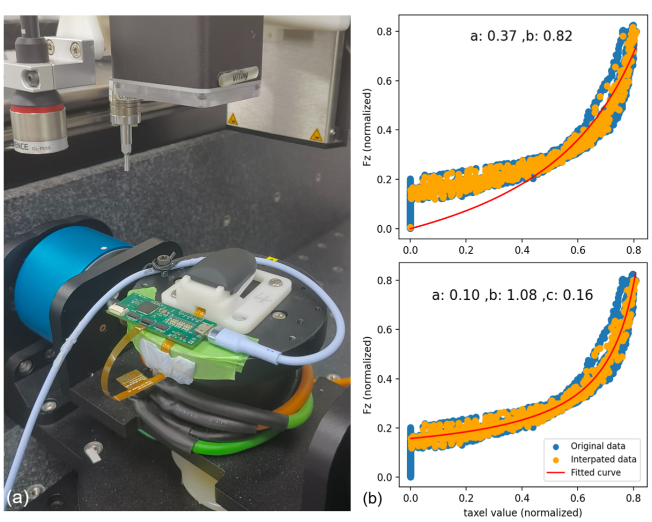

Fig. A1. Tactile sensor calibration: (a) setup for calibration, (b) calibration results for taxel 71 without (top) and with (bottom) the dead zone.

2) CMA-ES Optimization and Results: To improve the alignment between the simulated and real-world joint responses, the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [38] is employed. The optimization minimizes discrepancies in joint angles by adjusting the position control gain $K_p^{(sim)}$ and joint damping $K_d^{(sim)}$ for the simulated system. The optimization problem is formulated as:

$$\min_{K_p^{(sim)}, K_d^{(sim)}} \sum_{t=0}^{T} \|\theta_{\text{sim},t} - \theta_{\text{real},t}\|^2, \quad \text{s.t. } K_p^{(sim)}, K_d^{(sim)} > 0. \tag{A3}$$

where θ_{sim}, θ_{real} denote joint position sequences from simulation and real-world environments, respectively, with T representing the sequence length. The objective is to minimize the error between the joint angles in the simulation and the real system. TABLE A3 presents the initial CMA-ES parameters, while TABLE A4 lists the optimized joint parameters.

TABLE A3
CMA-ES Initial Parameters

Initial Guess	Step Size	Iterations	Tol	Lower Bounds	Upper Bounds
$[25.0, 1.0]$	0.5	30	0.05	$[0.01, 0.0]$	$[50.0, 5.0]$

TABLE A4
Optimized Joint Parameters

Joint	Position control gain $K_p^{(sim)}$	Joint damping $K_d^{(sim)}$
$F0_{\text{MPP}}$	26.27	2.37
$F0_{\text{DIP}}$	24.27	2.00
$F1_{\text{MPP}}$	49.32	3.96
$F1_{\text{DIP}}$	25.25	2.00
$F2_{\text{MPP}}$	49.92	3.59
$F2_{\text{DIP}}$	24.51	1.87

3) Domain Randomization Parameters: This section provides the detailed parameter ranges used during the domain randomization process, which are sampled during training to account for natural variations in object and environmental properties. The parameters are listed in TABLE A5. For example, the pen's mass ranges from 0.012 kg to 0.032 kg, and its diameter varies between 0.003 m and 0.006 m. Initial grasping positions are randomized along both the longitudinal and vertical axes to simulate real-world variations in grip. The starting orientation is set to $[0, 0, -1]$ to model the typical orientation during interaction. Friction coefficients are also randomized based on empirical data for various materials and surfaces. These ranges are adjusted to ensure realistic simulations that capture the wide range of interactions observed in real-world scenarios.

TABLE A5
Parameter Ranges for Domain Randomization

Parameter	**Range**
Pen mass (kg)	0.012∼0.032
Pen diameter (m)	0.003∼0.006
Initial grasp positions (m)	-0.006∼0.004 (longitudinal)
	0.0∼0.015 (vertical)
Starting orientation	[0, 0, -1]
Fingertip static friction coefficient	0.5∼1.0
Fingertip sliding friction coefficient	0.3∼0.7
Pen-tip friction coefficient	0.06∼0.3