CORTICALLY INSPIRED VISUO-TACTILE ALIGNMENT FOR CROSS-EMBODIMENT TACTILE INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Observing touch on another person's body can recruit somatosensory sensations in observers, providing a striking example of how the brain translates visual events into bodily states. Neuroscientific evidence suggests that this visuo-tactile resonance relies on structured alignment between visual and somatosensory neural representations, yet the computational instantiation and functional role of such alignment remain unclear. Here we introduce MTNet, a cortically inspired dual-stream framework that learns visual-to-tactile prediction under semantic, distributional, and geometric alignment constraints. Given monocular RGB observations, MTNet predicts force fields across 1,140 taxels on a robotic hand at 1-mm spatial resolution. Compared with reconstruction-driven models, the alignment constraints substantially improve spatiotemporal prediction and reshape visual representations to match the tactile manifold geometry, thereby simplifying the mapping from vision to touch. By aligning paired visual observations of human and robotic hands within the learned visuo-tactile space, the physical robot generates corresponding tactile states from observed human touch, achieving cross-domain transfer without human tactile labels. These results offer a computational account of how the cross-modal alignment supports visuo-tactile resonance across embodiments, translating neuroscientific findings into a design principle for tactile inference in embodied artificial systems. Code and videos are available in the supplementary materials and will be publicly released upon acceptance.

1 INTRODUCTION

Mirror touch refers to the phenomenon where observing touch on another person's body elicits corresponding tactile sensations in the observer (Serino et al., 2009; Ward & Banissy, 2015), exemplifying how perception can be transformed into bodily representation and supporting empathy (Lee Masson & Isik, 2023) and social cognition (Ward et al., 2018). This visuo-tactile resonance is not an incidental feature of sensory processing but reflects a fundamental organizing principle governing the integration of visual and somatosensory information in the brain (Schütz-Bosbach & Prinz, 2007). Therefore, characterizing this neural mechanism is a core question for both neuroscience and artificial intelligence, with implications for understanding cross-modal inference.

Neuroimaging studies have revealed that visuo-tactile resonance in the human brain relies on functional coupling between visual and somatosensory cortices. Observing touch co-activates these regions (Keysers et al., 2004; 2010), and crucially, a structural alignment exists between visual and somatosensory representations (Hedger et al., 2026): body-position tuning is systematically related to visual-field tuning and predicts preferences for both visual-field locations and body-part categories (Fig. 1a-b). This cortical alignment suggests that visuo-tactile resonance is grounded in a highly constrained mapping, potentially reflecting a structured computational principle. However, how to formalize this principle and verify whether it embodies a general computational rule for cross-modal integration remains an open question. Here we ask: Can cortical alignment principles be translated into a computational framework that governs visuo-tactile resonance?

To address this question, we propose Mirror Touch Net (MTNet) and its extension Adaptive MTNet (AMTNet) for learning aligned visuo-tactile representations and tactile prediction, validated on a dexterous robotic hand platform. Our key contributions are as follows:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

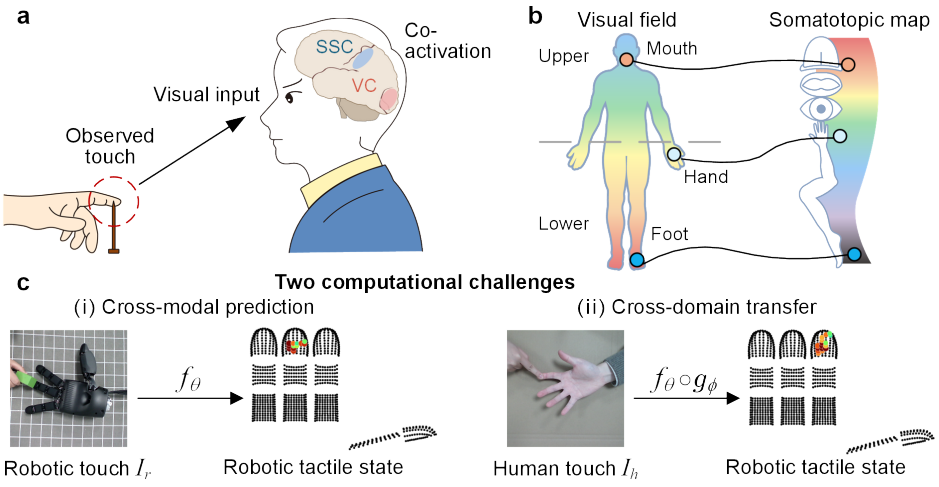


Figure 1: **Cortical visual-tactile resonance and computational challenges. a,** Observed touch on another body may co-activate the observer's visual cortex (VC) and somatosensory cortex (SSC), inducing a vicarious sensory response. **b,** This visuo-tactile resonance arises from aligned visual and somatosensory maps (adapted from Hedger et al. (2026)). **c,** We formulate this biological phenomenon as two progressive computational challenges on an embodied platform.

- **Cortically inspired architecture:** Inspired by the parietal visuo-tactile circuit, MTNet independently encodes visual and tactile inputs via separate pathways, projects them into a shared latent space, and decodes 3D force vectors across 1,140 discrete taxels from monocular RGB image sequences using a generative decoder.
- **Multi-level alignment constraints:** MTNet jointly optimizes five reconstruction losses and six additional alignment constraints that enforce distributional, semantic, and geometric consistency between visual and tactile modalities, embodying the cortical alignment principle and going beyond conventional reconstruction-driven objectives.
- **Label-free transfer:** Inheriting MTNet's pretrained parameters, AMTNet aligns novel visual representations of human hands into the learned visuo-tactile latent space without requiring human tactile labels, enabling observed human touch to generate corresponding robotic tactile states across different morphologies.

Experimental results on a robotic dexterous hand demonstrate that alignment constraints simplify cross-modal mapping. From monocular RGB input, MTNet predicts dense tactile signals at 1-mm resolution, outperforming reconstruction-driven baselines across five tailored metrics in both spatial and temporal accuracy. The aligned visuo-tactile space further enables cross-domain transfer without tactile supervision, where camera-observed human hand touch elicits corresponding tactile activations and reflexive motor responses in the physical robotic hand, reproducing the core phenomenology of mirror touch. These results provide a computational realization of cortical alignment principles in embodied systems, highlighting their functional role in cross-modal inference.

2 Related Work

Biological visuo-tactile resonance: Visuo-tactile resonance in the human brain, as exemplified by mirror touch, is a vital subject in neuroscience and cognitive science for studying multimodal integration. Neuroimaging studies have shown that observing touch activates cortical regions overlapping with those engaged by actual touch, primarily the secondary somatosensory cortex and occasionally the primary somatosensory cortex (Keysers et al., 2004; 2010). Visual and tactile signals converge in the parietal cortex, where they are integrated into shared topographic maps linking external visual space with internal bodily coordinates (Sereno & Huang, 2006; Pasalar et al., 2010; Klautke et al., 2023). Recent cross-modal mapping studies (Shahzad et al., 2025; Hedger et al., 2026) provide direct evidence for structural alignment between visual and somatosensory body rep-

resentations. However, this cortical alignment principle largely remains at the level of raw biological data, without computational modeling or exploration of its potential applications.

Robotic tactile sensing: The growing recognition of tactile sensing for dexterous manipulation has driven the development of diverse tactile sensing modalities, including electromagnetic (Yan et al., 2021), capacitive (Xu et al., 2024), and vision-based (Lambeta et al., 2020; Li et al., 2026a) approaches. These sensors have been integrated into embodied robots, particularly anthropomorphic hands (Sankar et al., 2025; Zhao et al., 2025; Zhang et al., 2025), to enable fine-grained contact estimation. However, the dominant paradigm usually treats touch primarily as a passive feedback signal: vision guides action selection, while tactile feedback refines motor commands (Chen et al., 2023; Huang et al., 2025; Bi et al., 2025; Li et al., 2026b). Tactile signals are available only during physical contact, making it a key challenge to enable robots to perform predictive tactile inference from visual observations, a capability that humans demonstrate routinely.

Visual-to-tactile prediction: Recent studies employed masked autoencoders (Higuera et al., 2025; Feng et al., 2025; Ye et al., 2026) and diffusion models (Zheng et al., 2026) for visuo-tactile integration and future tactile prediction, improving performance across various robotic tasks. However, these methods are primarily optimized by minimizing reconstruction losses without enforcing strong cross-modal consistency. Many existing approaches rely on vision-based tactile grippers (Zheng et al., 2026; Bi et al., 2025; Huang et al., 2025; Higuera et al., 2025; Feng et al., 2025; George et al., 2025) that treat both visual and tactile signals as image-like domains, yet despite their effectiveness, they do not fully reflect the human-like modality gap between high-dimensional, context-rich vision and low-dimensional, spatially sparse touch. Consequently, achieving cross-modal integration and prediction under a more pronounced visuo-tactile modality gap, alongside investigating the computational role of cortical alignment in this setting, remains an open problem.

3 METHOD

3.1 PROBLEM FORMULATION

We formulate biological visuo-tactile resonance as two progressive computational challenges on an embodied system (Fig. 1c): **(i)** predicting self-tactile state from vision; **(ii)** transferring this prediction to human touch without tactile labels. Solving these two challenges provides a computable and verifiable platform for studying how visual observations can evoke embodied tactile sensation. Our platform employs a robotic hand with 11 tactile sensors across fingertips and fingerpads, totaling 1,140 independent taxels (Fig. S1). Each taxel records three-axis forces at 1-mm spatial resolution, comparable to human skin (Johnson & Phillips, 1981), and is represented by a six-dimensional vector $[x, y, z, f_x, f_y, f_z]$ comprising spatial coordinates and force components. Visual input comes from a monocular RGB camera at 30 Hz, resized to 224×224 pixels. This setup reflects the sparse, high-resolution nature of human tactile sensing versus context-rich visual frames.

Formally, let $I_r \in \mathbb{R}^{3K\times224\times224}$ be a stack of K consecutive RGB frames, and $O \in \mathbb{R}^{P\times N\times3}$ the robotic ground-truth tactile force field at the K-th time step, where $P = 11$ is the number of sensing regions and $N = 120$ is the maximum number of taxels per sensor. A binary mask $M \in \{0, 1\}^{P\times N}$ indicates valid taxels. The cross-modal prediction learns $f_\theta : I_r \mapsto \hat{O}_v$ to minimize the error between the predicted field $\hat{O}_v$ and O. For cross-domain transfer, let I_h denote human-hand frames of similar contacts; the goal is to learn a composite mapping $f_\theta \circ g_\phi : I_h \mapsto \hat{O}_h$, where g_ϕ is the domain adapter and $\hat{O}_h$ the robotic tactile fields generated from human visual input. We address both challenges through the cortically inspired architecture and constraints described below.

3.2 MTNET ARCHITECTURE

MTNet learns $f_\theta : I_r \mapsto \hat{O}_v$ to infer 3D tactile force fields from visual observations. The architecture is motivated by the parallel processing of visual and somatosensory neural circuits and their subsequent integration in the parietal cortex (Sereno & Huang, 2006; Pasalar et al., 2010; Klautke et al., 2023). As illustrated in Fig. 2a-b, separate visual and tactile pathways produce region-level representations that are aligned in a shared latent space through cross-modal consistency objectives. A conditioned decoder shared by both pathways then generates taxel-level force fields. The tactile pathway provides supervisory targets during training; inference requires visual input alone.

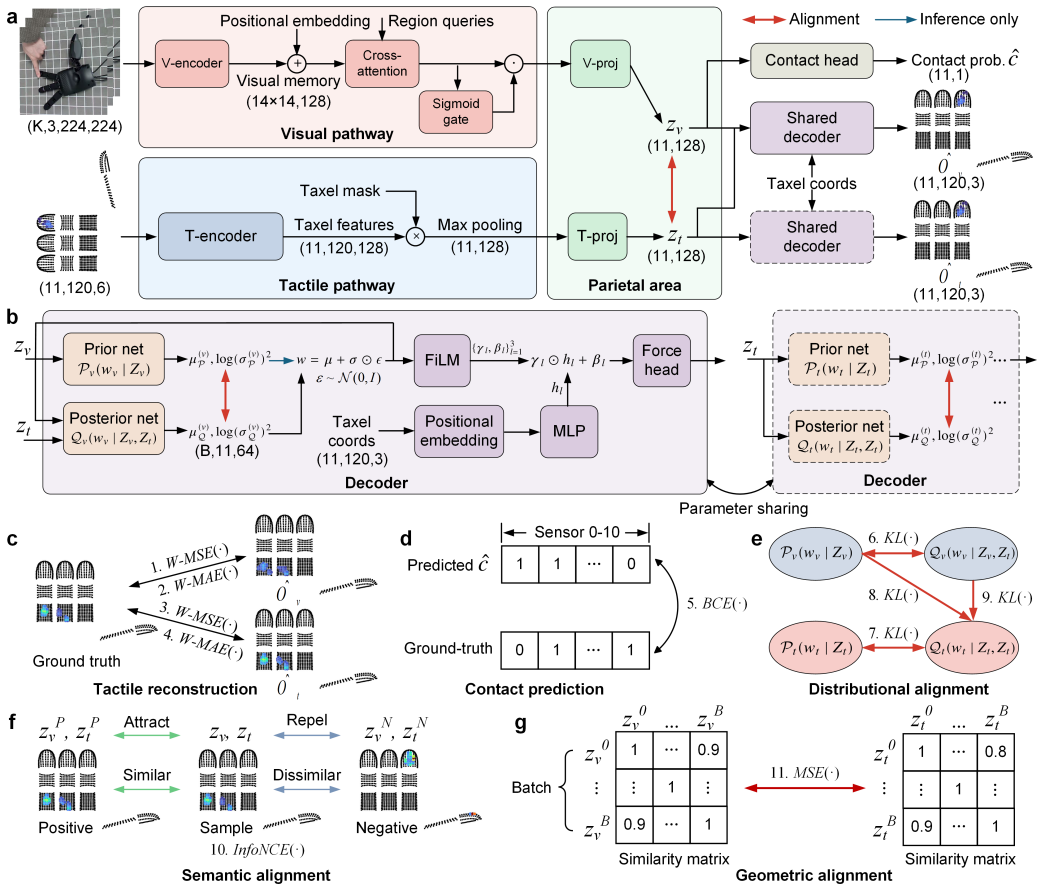


Figure 2: **MTNet architecture and training objectives. a,** Separate visual and tactile pathways project their inputs into a shared parietal-like latent space, followed by a common tactile decoder. **b,** Paired features condition the latent posterior during training, whereas inference relies solely on the visual prior; tactile self-reconstruction anchors the shared decoder. **c–g,** Eleven training objectives.

Visual pathway. The visual pathway produces region-level visual tokens Z_v and predicts per-region contact probabilities $\hat{c}$ (Fig. 2a). From the input stack I_r, each frame is encoded into patch features by a modified ResNet-18 backbone (He et al., 2016). A lightweight temporal attention module aggregates these features across frames, and the resulting map is augmented with patch positional embeddings to form the visual memory $V \in \mathbb{R}^{S\times d}$, where S is the number of spatial patches and $d = 128$ the latent feature dimension. Computational details are provided in Appendix A.1.

Then, learnable sensor queries $Q \in \mathbb{R}^{P\times d}$ retrieve region-specific evidence from visual memory V through multi-head cross-attention (MHCA) (Vaswani et al., 2017), yielding attended features $H = \mathrm{MHCA}(Q, V, V) \in \mathbb{R}^{P\times d}$. Motivated by inhibitory modulation in biological somatosensory processing (Fan et al., 2020), a feature-wise gate suppresses irrelevant features. After gating, a projection head maps the filtered features into a latent space aligned with tactile representations, producing region-level visual tokens $Z_v \in \mathbb{R}^{P\times d}$. A linear classifier subsequently predicts the contact probability for each region:

$$Z_v = \Pi_v(H \odot \sigma(\mathcal{G}_v(H))) \in \mathbb{R}^{P\times d}, \quad \hat{c} = \sigma(\mathcal{C}_v(Z_v)) \in [0,1]^P \tag{1}$$

, where $\mathcal{G}_v$, Π_v, and $\mathcal{C}_v$ denote the gating network, projection head, and contact classifier, respectively; σ is the sigmoid function, and $\odot$ denotes element-wise multiplication. The resulting $\hat{c}$ is used to modulate the corresponding taxel-level force predictions in Eq. 5.

Tactile pathway. The tactile pathway encodes measured tactile states into region-level tokens Z_t, with which visual tokens Z_v are aligned during training. We denote the tactile input by $I^t \in \mathbb{R}^{P\times N\times 6}$, where each taxel n in sensing region p is represented by $I^t_{p,n} = [x, y, z, f_x, f_y, f_z]$,

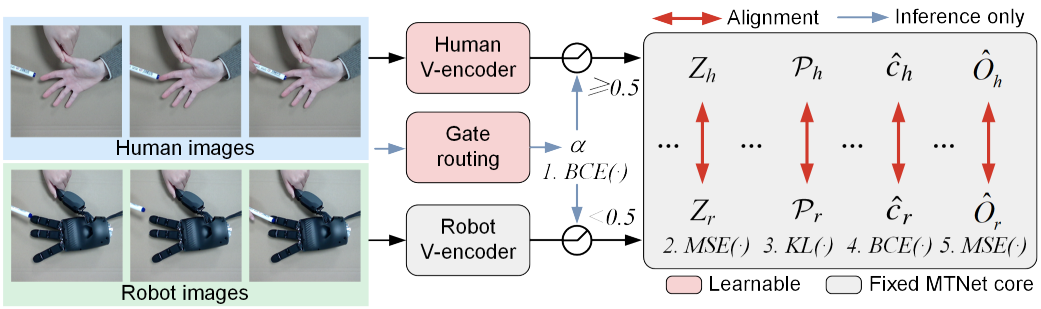


Figure 3: **AMTNet for cross-domain transfer of tactile prediction.** By aligning paired observations, AMTNet maps human-hand images into MTNet's frozen visuo-tactile space, thereby extending tactile prediction to the human domain without human tactile supervision.

comprising its spatial coordinates and measured force vector. We employ a grouped PointNet-based encoder (Qi et al., 2017) to process each region independently with shared parameters. Taxel-wise feature lifting, masked max pooling, and latent projection are applied sequentially to obtain the region-level tactile tokens:

$$Z_{t,p} = \Pi_t\left(\underset{n:\, M_{p,n}=1}{\text{MaxPool}}\, \Phi\big(I^t_{p,n}\big)\right) \in \mathbb{R}^d, \qquad p = 1, \ldots, P, \tag{2}$$

where $\Phi : \mathbb{R}^6 \to \mathbb{R}^d$ denotes the shared taxel-wise encoder with $d = 128$; Π_t denotes a two-layer projection head. The binary mask $M \in \{0,1\}^{P\times N}$ ensures that padded entries are excluded from the pooling operation. Stacking the tokens from all sensing regions yields $Z_t \in \mathbb{R}^{P\times d}$, matching the region-level organization of Z_v for subsequent cross-modal alignment and tactile decoding.

Generative decoding. Given the region-level representations Z_v and Z_t, MTNet uses a conditional variational autoencoder (CVAE) (Sohn et al., 2015) to generate taxel-level force fields. During training, the decoder is applied to both modalities: conditioning on Z_v yields the visual-to-tactile prediction, whereas conditioning on Z_t yields an auxiliary tactile self-reconstruction. The two branches share all decoder parameters.

To capture the uncertainty of force distributions, we introduce a region-level stochastic latent variable $w_m \in \mathbb{R}^{P\times d_w}$, where $d_w = 64$ and $m \in \{v, t\}$ indexes the visual and tactile conditioning modality. For each branch, the shared prior and posterior networks parameterize two diagonal Gaussian distributions, from which w_m is sampled using the reparameterization trick:

$$\begin{aligned} &[\mu^{(m)}_{\mathcal{P}}, \ell^{(m)}_{\mathcal{P}}] = \Phi_{\mathcal{P}}(Z_m), \quad [\mu^{(m)}_{\mathcal{Q}}, \ell^{(m)}_{\mathcal{Q}}] = \Phi_{\mathcal{Q}}([Z_m, Z_t]), \\ &w_m = \mu^{(m)}_{\mathcal{S}} + \epsilon \odot \exp\left(\tfrac{1}{2}\ell^{(m)}_{\mathcal{S}}\right), \epsilon \sim \mathcal{N}(0, I) \end{aligned} \tag{3}$$

, where $\ell^{(m)}_{\mathcal{S}} = \log(\sigma^{(m)}_{\mathcal{S}})^2$ for $\mathcal{S} \in \{\mathcal{P}, \mathcal{Q}\}$, and $[\cdot,\cdot]$ denotes feature concatenation. The first line parameterizes the conditional prior $\mathcal{P}_m(w_m \mid Z_m)$ and approximate posterior $\mathcal{Q}_m(w_m \mid Z_m, Z_t)$, respectively. During training, $\mathcal{S} = \mathcal{Q}$ and both branches sample from their posteriors. At inference, only the visual branch is evaluated, with $m = v$ and $\mathcal{S} = \mathcal{P}$, thereby requiring no tactile input.

The sampled latent variable and deterministic representations condition an implicit neural field that decodes taxel coordinates into three-axis forces. Coordinates are first encoded with multi-frequency Fourier features, then queried against a shared FiLM-conditioned residual MLP (Perez et al., 2018):

$$\begin{aligned} E_c(C) &= \left[C,\, \left\{\sin(2^j\pi C), \cos(2^j\pi C)\right\}_{j=0}^{J-1}\right], \qquad J = 10, \\ \hat{O}^{\text{raw}}_m &= \mathcal{D}^{\text{FiLM}}_\theta\left(E_c(C); \Phi_{\text{film}}([w_m, Z_m])\right), \qquad m \in \{v, t\} \end{aligned} \tag{4}$$

, where $C \in \mathbb{R}^{P\times N\times 3}$ denotes the taxel coordinates, $E_c(C) \in \mathbb{R}^{P\times N\times d_c}$ is their Fourier encoding, and Φ_{film} generates region-level modulation parameters from the stochastic latent variable w_m and deterministic condition Z_m. The shared decoder $\mathcal{D}^{\text{FiLM}}_\theta$ comprises three FiLM-modulated residual layers, with each region-level modulation broadcast over its taxels, and outputs $\hat{O}^{\text{raw}}_m \in \mathbb{R}^{P\times N\times 3}$.

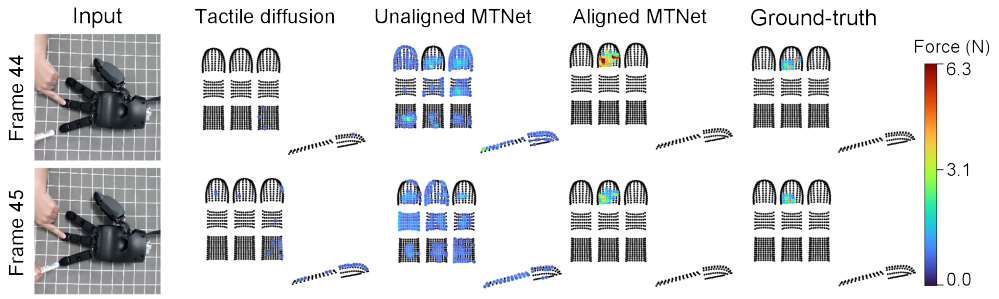


Figure 4: **Tactile prediction snapshots.** Each panel visualizes the predicted 3D force vectors for all 1140 taxels, with colors encoding the resultant force magnitude (see Supplementary Video 1).

Finally, the region-level contact probabilities $\hat{c}$ predicted by the visual pathway (Eq. 1) provide a contact-conditioned gain for the visual output:

$$\hat{O}_v = \hat{O}_v^{\text{raw}} \odot (1 + \text{Expand}(\hat{c})), \qquad \hat{O}_t = \hat{O}_t^{\text{raw}}. \tag{5}$$

Here, $\text{Expand}(\hat{c})$ broadcasts the contact probabilities across the taxel and force dimensions. The final visual output $\hat{O}_v$ is MTNet's visual-to-tactile prediction, while tactile self-reconstruction via $\hat{O}_t$ grounds the tactile posterior as an anchor for the cross-modal alignment introduced next.

3.3 Optimization objectives

The core idea of MTNet is to jointly optimize tactile prediction and cross-modal consistency. Beyond minimizing reconstruction errors, it enforces alignment across distributional, semantic, and geometric levels. The overall objective comprises five reconstruction losses and six alignment terms.

Reconstruction terms. As shown in Fig. 2c-d, the decoded outputs $\hat{O}_v$ and $\hat{O}_t$ (Eq. 5) are optimized against the ground-truth force field O via weighted masked mean squared errors (W-MSE) $\mathcal{L}_{\text{rec}}^m$ on force components and mean absolute errors (W-MAE) $\mathcal{L}_{\text{mag}}^m$ on magnitudes. To account for sparse contact, both losses upweight valid taxels carrying nonzero ground-truth force. The predicted contact probabilities $\hat{c}$ (Eq. 1) are supervised by a class-balanced binary cross-entropy (BCE) loss $\mathcal{L}_{\text{ct}}$, using region-level labels derived from O. Detailed formulations are provided in Appendix A.2.1.

Distributional alignment. Four Kullback–Leibler (KL) constraints regularize the conditional latent distributions (Fig. 2e). Beyond the standard CVAE terms $\mathcal{L}_{\text{KL}}^m = D_{\text{KL}}(\mathcal{Q}_m \| \mathcal{P}_m)$, $m \in \{v, t\}$, we introduce two cross-modal losses $\mathcal{L}_{\text{KL}}^{Q_v Q_t} = D_{\text{KL}}(\mathcal{Q}_v \| \bar{\mathcal{Q}}_t)$ and $\mathcal{L}_{\text{KL}}^{P_v Q_t} = D_{\text{KL}}(\mathcal{P}_v \| \bar{\mathcal{Q}}_t)$, which align the visual posterior and prior, respectively, with the EMA-smoothed tactile posterior $\bar{\mathcal{Q}}_t$. The measurement-grounded posterior anchors the transfer of tactile distributional structure to both the visual posterior during training and the visual prior during inference. The EMA update and KL-weight warm-up are detailed in Appendix A.2.2.

Semantic alignment. Deterministic visual and tactile tokens (Z_v and Z_t) are also aligned using a region-wise InfoNCE objective (Oord et al., 2018). We define a positive distribution Π_Y from the cosine similarity of force fields, thereby including samples with similar force patterns as positives (Fig. 2f). Let $\mathcal{P}_Y$ denote the positive pairs encoded by Π_Y and $\mathcal{N}_Y$ their complementary negative pairs. The force-aware contrastive loss is:

$$\mathcal{L}_{\text{con}} = \frac{1}{2P} \sum_{p=1}^{P} [\ell_{\text{NCE}}(Z_v^p, Z_t^p; \mathcal{P}_Y, \mathcal{N}_Y, \Pi_Y) + \ell_{\text{NCE}}(Z_t^p, Z_v^p; \mathcal{P}_Y, \mathcal{N}_Y, \Pi_Y)] \tag{6}$$

, where ℓ_{NCE} attracts force-similar positive pairs according to Π_Y while contrasting them against negative pairs. This objective preserves region correspondence while promoting physically consistent cross-modal representations; implementation details are provided in Appendix A.2.3.

Geometric alignment. We introduce a relational loss $\mathcal{L}_{\text{rel}}$ that enforces consistent sample-to-sample geometry between modalities (Fig. 2g). It minimizes the mean-squared difference between the pairwise similarity structures of normalized sample-level descriptors from Z_v and Z_t, encouraging them to preserve similar inter-sample topology in the shared latent space. See Appendix A.2.4 for details.

Table 1: **Ablation of alignment constraints in MTNet.** Results are means over five runs. "w/o" indicates removal of the specified constraint. ↑: higher is better; ↓: lower is better.

Models	NRMSE ↓	S-CosSim ↑	S-CCC ↑	T-IoU ↑	W-F1 ↑
Tactile diffusion (Zheng et al., 2026)	0.0512	0.0629	0.7391	0.0753	0.1310
Unaligned MTNet	0.0576	0.1682	0.0090	0.0816	0.1397
MTNet w/o distributional alignment	0.0640	0.2217	0.0404	0.0818	0.1397
MTNet w/o semantic alignment	0.0101	0.5921	0.8669	0.7364	0.9524
MTNet w/o geometric alignment	**0.0098**	0.6078	0.8710	0.7573	0.9553
Aligned MTNet	0.0102	**0.6097**	**0.8814**	**0.7920**	**0.9664**

Thus, the total objective of MTNet is a weighted sum of individual loss terms with weights λ:

$$\begin{aligned}\mathcal{L}^{\mathrm{M}}_{\mathrm{total}} = &\sum_{m\in\{v,t\}} \left(\lambda^{m}_{\mathrm{rec}}\mathcal{L}^{m}_{\mathrm{rec}} + \lambda^{m}_{\mathrm{mag}}\mathcal{L}^{m}_{\mathrm{mag}}\right) + \lambda_{\mathrm{ct}}\mathcal{L}_{\mathrm{ct}} + \lambda_{\mathrm{KL}}\left(\mathcal{L}^{v}_{\mathrm{KL}} + \mathcal{L}^{t}_{\mathrm{KL}}\right) \\ &+ \lambda_{\mathcal{Q}_v\mathcal{Q}_t}\mathcal{L}^{\mathcal{Q}_v\mathcal{Q}_t}_{\mathrm{KL}} + \lambda_{\mathcal{P}_v\mathcal{Q}_t}\mathcal{L}^{\mathcal{P}_v\mathcal{Q}_t}_{\mathrm{KL}} + \lambda_{\mathrm{con}}\mathcal{L}_{\mathrm{con}} + \lambda_{\mathrm{rel}}\mathcal{L}_{\mathrm{rel}}\end{aligned} \tag{7}$$

3.4 AMTNet for cross-domain transfer

Next, we extend MTNet's cross-modal alignment to bridge the morphological gap between human and robotic hands, enabling tactile prediction from human visual observations. A key challenge is the absence of tactile labels for human hands, which prevents direct training with paired visual-tactile data as in Sec. 3.3. We therefore introduce AMTNet (Fig. 3), a cross-domain transfer strategy that aligns human visual space with the robot visual space learned by MTNet. This alignment indirectly links to the robot tactile space through MTNet's pretrained visual-tactile mapping, using paired human-hand images whose contact configurations match those of the robotic hand.

AMTNet preserves MTNet's visual-tactile alignment by freezing its pretrained robot visual modules, contact head, and tactile decoder. It adds only a human-specific encoder $\mathcal{E}_h$ which is initialized from the robot visual encoder $\mathcal{E}_r$, and a lightweight gate G_ϕ; these are the only trainable components during cross-domain adaptation. Given a paired human–robot image pair (I_h, I_r) with domain labels $y = 1$ for human and $y = 0$ for robot, the two encoders produce $U_h = \mathcal{E}_h(I_h)$ and $U_r = \mathcal{E}_r(I_r)$. The gate predicts the human-domain probability $\alpha = G_\phi(I)$ and performs soft routing:

$$U_{\mathrm{mix}} = \alpha U_h + (1-\alpha)U_r, \quad \mathcal{L}_{\mathrm{gate}} = \mathbb{E}_{(I,y)}\left[-y\log\alpha - (1-y)\log(1-\alpha)\right] \tag{8}$$

, encouraging $\alpha \to 1$ for human images and $\alpha \to 0$ for robot images. At inference, hard routing selects $\mathcal{E}_h$ if $\alpha \geq 0.5$ else $\mathcal{E}_r$. During training, the mixed feature U_{mix} passes through the frozen MTNet core to produce student outputs $(Z_h, \mathcal{P}_h, \hat{c}_h, \hat{O}_h)$, while the robot encoder directly processes I_r to yield detached teacher outputs $(Z_r, \mathcal{P}_r, \hat{c}_r, \hat{O}_r)$. We align the visual representations, prior distributions, and tactile predictions between the two streams using $\mathcal{L}^{\mathrm{A}}_{\mathrm{feat}}$, $\mathcal{L}^{\mathrm{A}}_{\mathrm{KL}}$, $\mathcal{L}^{\mathrm{A}}_{\mathrm{ct}}$, and $\mathcal{L}^{\mathrm{A}}_{\mathrm{pred}}$, together with the gate loss. Detailed formulations are provided in Appendix A.3. The total objective is:

$$\mathcal{L}^{\mathrm{A}}_{\mathrm{total}} = \lambda^{\mathrm{A}}_{\mathrm{feat}}\mathcal{L}^{\mathrm{A}}_{\mathrm{feat}} + \lambda^{\mathrm{A}}_{\mathrm{pred}}\mathcal{L}^{\mathrm{A}}_{\mathrm{pred}} + \lambda^{\mathrm{A}}_{\mathrm{ct}}\mathcal{L}^{\mathrm{A}}_{\mathrm{ct}} + \lambda^{\mathrm{A}}_{\mathrm{KL}}\mathcal{L}^{\mathrm{A}}_{\mathrm{KL}} + \lambda^{\mathrm{A}}_{\mathrm{gate}}\mathcal{L}^{\mathrm{A}}_{\mathrm{gate}} \tag{9}$$

, where λ^{A} denotes the loss weight. By updating only the human encoder and gate, AMTNet maps human observations into the established robot visual-tactile space without human tactile supervision, enabling the frozen MTNet decoder to predict tactile force fields directly from human-hand images.

4 Experimental results

We conduct comprehensive experiments to validate the proposed framework, covering tactile prediction performance, the induced manifold analysis, and cross-domain adaptation to human-hand touch. Dataset, implementation, and metric details are provided in Appendix A.4–A.6.

4.1 Visual-to-tactile prediction through alignment

We first evaluate MTNet on predicting tactile fields from contact images of a robotic hand. It is compared against a recent vision-conditioned tactile diffusion model (Zheng et al., 2026) that pre-

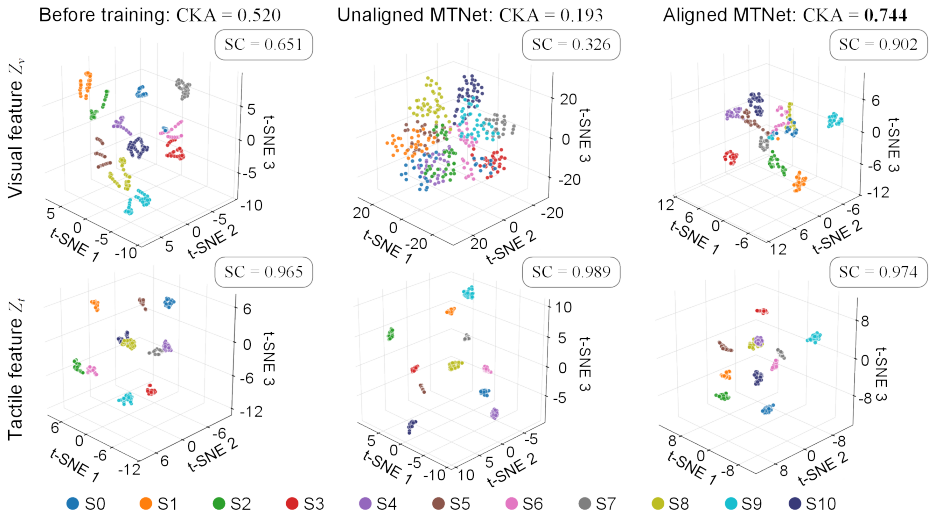


Figure 5: **t-SNE of visual and tactile features from isolated touch events of 11 sensors (S0-S10).**

dicts tactile responses via a denoising objective but does not enforce explicit cross-modal alignment. Five complementary metrics are used to assess numerical errors, spatial distribution, temporal consistency, and contact timing; baseline implementation and metric computation are detailed in Appendix A.5–A.6. Despite visual occlusion, MTNet reliably infers contact location, timing, and force magnitude, outperforming the baseline across all metrics (Fig. 4 and Tab. 1). The diffusion baseline performs acceptably on the temporal-dynamic metric but fails to reconstruct spatial contact patterns.

We further ablate the contribution of each alignment constraint. Removing any individual constraint degrades performance relative to the fully aligned MTNet (Tab. 1). Notably, both baselines exhibit normally decreasing training losses (Fig. S4–S5), ruling out optimization failure as the cause of poor prediction. Rather, since many temporal and spatial locations may correspond to non-contact states, reconstruction losses can be reduced by fitting zero values and mean responses. Thus, alignment constraints provide essential cross-modal supervision that anchors the optimization objective to contact-relevant structures. These results confirm that pure numerical regression is insufficient for visuo-to-tactile prediction and that alignment constraints directly improve tactile prediction, as further supported by manifold analysis in the following section.

4.2 Manifold Analysis of Cross-Modal Alignment

To explain how alignment improves tactile prediction, we perform a manifold analysis using t-SNE (Van der Maaten & Hinton, 2008). We compute the latent manifolds of visual and tactile features from isolated contact events, where each event independently activates one of the 11 sensors (Fig. 5). Intrinsic tactile manifold naturally forms well-separated clusters corresponding to distinct contact events, whereas shared backgrounds and subtle visual variations blur the visual distinctions, resulting in an initial visual silhouette score (SC) (Rousseeuw, 1987) approximately 0.31 below the tactile score. This organizational discrepancy indicates the difficulty of mapping visual inputs to a structured tactile space. Reconstruction-only training further widens this gap to 0.66, presumably because direct regression favors averaged responses that minimize numerical error, making distinct touch events less separable in the visual latent space and thereby limiting predictive capacity.

In contrast, the fully aligned MTNet reorganizes the visual manifold to be consistent with tactile geometry, achieving the highest centered kernel alignment (CKA) similarity (Kornblith et al., 2019) of 0.74, exceeding the untrained and unaligned models by 0.22 and 0.55, respectively. Pairwise cosine similarity corroborates these findings (Fig. S6): the untrained and unaligned models produce largely uniform visual similarity matrices with weak cross-modal correspondence, while the full MTNet recovers clear structure and diagonal alignment. These results show that numerical regression alone cannot bridge the inherent structural discrepancy between vision and touch, and that alignment constraints impose structural consistency, thereby simplifying the mapping to tactile space.

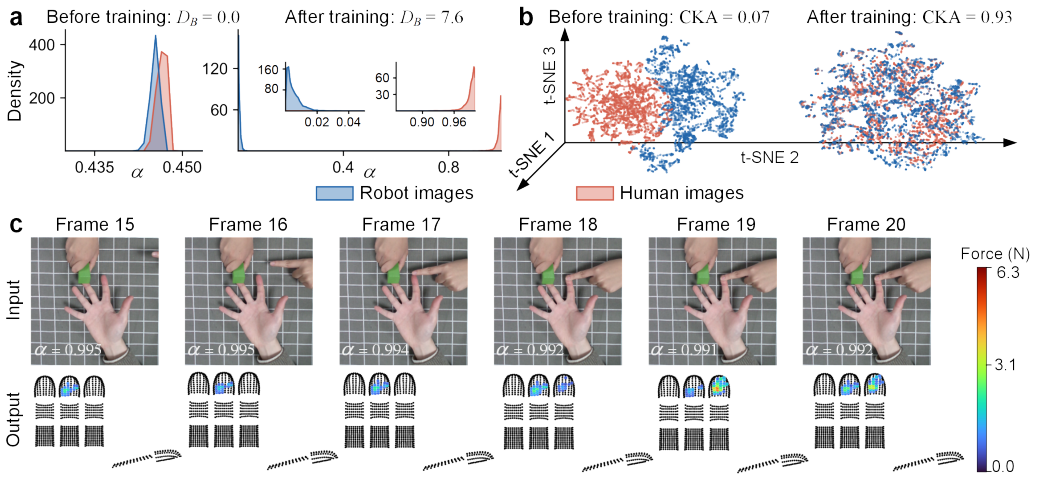


Figure 6: **AMTNet for vicarious tactile prediction. a,** Gating output α distribution ($\alpha \to 1$ human, $\to 0$ robot). **b,** t-SNE of visual features Z_v from human and robot images. **c,** AMTNet prediction snapshots on human-hand images (see Supplementary Video 2).

4.3 Transfer of Tactile Prediction to Human Hands

We next evaluate whether AMTNet transfers tactile prediction from robotic to human hands. The gating distributions (Eq. 8) for the two domains transition from overlapping to separated, with Bhattacharyya distance (D_B) (Bhattacharyya, 1943) increasing from 0 to 7.6; visual representations shift from distinct to converged, with CKA similarity rising from 0.07 to 0.93 (Fig. 6a-b). Through this alignment, AMTNet predicts robotic tactile signals from human-hand images, including contact locations and force magnitudes (Fig. 6c), matching those from corresponding robot images (Fig. S7) and thereby demonstrating cross-domain tactile prediction without human tactile supervision.

Finally, we test whether AMTNet elicits mirror-touch-like behavior on the physical robotic hand (Fig. S8). The system triggers a finger flick when the maximum contact force exceeds 0.2 N. In the main experiment, a human hand is touched under a camera while the robotic hand remains untouched. AMTNet predicts tactile signals from human-hand frames, and these predictions suffice to trigger flick responses in the corresponding robotic fingers. Predicted force trajectories closely align with visual motion traces, confirming synchronized responses to observed touch. The system also generalizes to objects unseen in the training datasets (Supplementary Video 3), suggesting the alignment process focuses on core contact events while suppressing irrelevant visual cues.

5 Conclusion

In this study, we show that cross-modal alignment serves as a critical computational principle for cross-embodiment tactile sensation. Inspired by cortical alignment, our framework MTNet and its adaptive extension AMTNet impose alignment constraints across semantic, distributional, and geometric levels, which simplifies cross-modal mapping from dense vision to sparse touch, thereby enabling label-free cross-domain transfer from robot to human hands. These results bridge neuroscientific findings of visuo-tactile resonance with embodied implementations, advancing our understanding of cross-modal integration in the brain and broadening its applications.

Limitations: Since contact events inherently cause occlusion, monocular RGB observations cannot fully recover precise 3D force values at the taxel level. The current system is therefore more reliable in estimating contact locations and coarse force magnitudes. Furthermore, our robotic hand is passively touched in the experiments, which isolates the role of tactile prediction but leaves closed-loop sensorimotor behavior unexplored. Future work should extend the alignment framework to diverse contact conditions and sensory modalities, and combine tactile inference with motor control to achieve more complex cross-embodiment generalization and human-robot interaction.

AI USE STATEMENT

In this work, we used generative AI tools for writing polish, proofreading spelling and grammar errors. We have not used generative AI tools for designing methods and experiments or interpreting results, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. The authors rigorously reviewed the LLM-polished text to ensure accuracy and adherence to standards. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

No ethical issues are associated with this work.

REPRODUCIBILITY STATEMENT

Code for MTNet, AMTNet, and the tactile diffusion baseline is provided in the supplementary materials and will be open-sourced upon acceptance. Three demonstration videos are also included to illustrate our method and experimental setup. Computational procedures and optimization objectives appear in Appendices A.1–A.3; dataset description, training details, and metric computations are in Appendices A.4–A.6. Together, these resources ensure the reproducibility of our work.

REFERENCES

Ayan Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. 1943. URL https://api.semanticscholar.org/CorpusID:235941388.

Jianxin Bi, Kevin Yuchen Ma, Ce Hao, Mike Zheng Shou, and Harold Soh. Vla-touch: Enhancing vision-language-action models with dual-level tactile feedback. *arXiv preprint arXiv:2507.17294*, 2025.

Yizhou Chen, Mark Van der Merwe, Andrea Sipos, and Nima Fazeli. Visuo-tactile transformers for manipulation. In *Conference on Robot Learning*, pp. 2026–2040. PMLR, 2023.

Linlin Z Fan, Simon Kheifets, Urs L Böhm, Hao Wu, Kiryl D Piatkevich, Michael E Xie, Vicente Parot, Yooree Ha, Kathryn E Evans, Edward S Boyden, et al. All-optical electrophysiology reveals the role of lateral inhibition in sensory processing in cortical layer 1. *Cell*, 180(3):521–535, 2020.

Ruoxuan Feng, Jiangyu Hu, Wenke Xia, Tianci Gao, Ao Shen, Yuhao Sun, Bin Fang, and Di Hu. Anytouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 31265–31285, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/4d893f766ab60e5337659b9e71883af4-Paper-Conference.pdf.

Abraham George, Selam Gano, Pranav Katragadda, and Amir Barati Farimani. Vital pretraining: Visuo-tactile pretraining for tactile and non-tactile manipulation policies. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 258–264. IEEE, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Nicholas Hedger, Thomas Naselaris, Kendrick Kay, Tomas Knapen, and Magdalena Skipper. Vicarious body maps bridge vision and touch in the human brain. *Nature*, 2026.

Carolina Higuera, Akash Sharma, Chaithanya Krishna Bodduluri, Taosha Fan, Patrick Lancaster, Mrinal Kalakrishnan, Michael Kaess, Byron Boots, Mike Lambeta, Tingfan Wu, et al. Sparsh: Self-supervised touch representations for vision-based tactile sensing. In *Conference on Robot Learning*, pp. 885–915. PMLR, 2025.

Binghao Huang, Yixuan Wang, Xinyi Yang, Yiyue Luo, and Yunzhu Li. 3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing. In *Conference on Robot Learning*, pp. 2557–2578. PMLR, 2025.

Kenneth O Johnson and John R Phillips. Tactile spatial resolution. i. two-point discrimination, gap detection, grating resolution, and letter recognition. *Journal of neurophysiology*, 46(6):1177–1192, 1981.

Christian Keysers, Bruno Wicker, Valeria Gazzola, Jean-Luc Anton, Leonardo Fogassi, and Vittorio Gallese. A touching sight: Sii/pv activation during the observation and experience of touch. *Neuron*, 42(2):335–346, 2004.

Christian Keysers, Jon H Kaas, and Valeria Gazzola. Somatosensation in social perception. *Nature Reviews Neuroscience*, 11(6):417–428, 2010.

Janina Klautke, Celia Foster, W Pieter Medendorp, and Tobias Heed. Dynamic spatial coding in parietal cortex mediates tactile-motor transformation. *Nature Communications*, 14(1):4532, 2023.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019.

Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020.

Haemy Lee Masson and Leyla Isik. Rapid processing of observed touch through social perceptual brain regions: an eeg-fmri fusion study. *The Journal of neuroscience*, 43(45):7700–7711, 2023.

Shoujie Li, Tong Wu, Jianle Xu, Yan Huang, Zongwen Zhang, Hongfa Zhao, Qinghao Xu, Zihan Wang, Linqi Ye, Yang Yang, et al. Biomimetic multimodal tactile sensing enables human-like robotic perception. *Nature Sensors*, 1(1):52–62, 2026a.

Yuyang Li, Yinghan Chen, Zihang Zhao, Puhao Li, Tengyu Liu, Siyuan Huang, and Yixin Zhu. Simultaneous tactile-visual perception for learning multimodal robot manipulation. *IEEE Robotics and Automation Letters*, 2026b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. URL https://api.semanticscholar.org/CorpusID:53592270.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Siavash Pasalar, Tony Ro, and Michael S. Beauchamp. Tms of posterior parietal cortex disrupts visual tactile multisensory integration. *European Journal of Neuroscience*, 31(10):1783–1790, 2010.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. FiLM: Visual Reasoning with a General Conditioning Layer. In *AAAI Conference on Artificial Intelligence*, pp. 3942–3951, 2018. doi: 10.1609/AAAI.V32I1.11671. URL https://mlanthology.org/aaai/2018/perez2018aaai-film/.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.

Peter Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1):53–65, November 1987. ISSN 0377-0427. doi: 10.1016/0377-0427(87)90125-7. URL https://doi.org/10.1016/0377-0427(87)90125-7.

Sriramana Sankar, Wen-Yu Cheng, Jinghua Zhang, Ariel Slepyan, Mark M Iskarous, Rebecca J Greene, Rene DeBrabander, Junjun Chen, Arnav Gupta, and Nitish V Thakor. A natural biomimetic prosthetic hand with neuromorphic tactile sensing for precise and compliant grasping. *Science Advances*, 11(10):eadr9300, 2025.

Simone Schütz-Bosbach and Wolfgang Prinz. Perceptual resonance: action-induced modulation of perception. *Trends in cognitive sciences*, 11 8:349–55, 2007. URL https://api.semanticscholar.org/CorpusID:43831896.

Martin I Sereno and Ruey Song Huang. A human parietal face area contains aligned head-centered visual and tactile maps. *Nature Neuroscience*, 9(10):1337–1343, 2006.

Andrea Serino, Giulia Giovagnoli, and Elisabetta Làdavas. I feel what you feel if you are similar to me. *PloS one*, 4(3):e4930, 2009.

Iqra Shahzad, Ceren Battal, Filippo Cerpelloni, Alice Van Audenhaege, André Mouraux, and Olivier Collignon. Aligned representation of visual and tactile motion directions in hmt+/v5 and fronto-parietal regions. *Nature Communications*, 17, 2025. URL https://api.semanticscholar.org/CorpusID:277317888.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jamie Ward and Michael J Banissy. Explaining mirror-touch synesthesia. *Cognitive neuroscience*, 6(2-3):118–133, 2015.

Jamie Ward, Patricia Schnakenberg, and Michael J Banissy. The relationship between mirror-touch synaesthesia and empathy: New evidence and a new screening tool. *Cognitive neuropsychology*, 35(5-6):314–332, 2018.

Boxin Xu, Luoyan Zhong, Grace Zhang, Xiaoyu Liang, Diego Virtue, Rishabh Madan, and Tapomayukh Bhattacharjee. Cushsense: Soft, stretchable, and comfortable tactile-sensing skin for physical human-robot interaction. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5694–5701. IEEE, 2024.

Youcan Yan, Zhe Hu, Zhengbao Yang, Wenzhen Yuan, Chaoyang Song, Jia Pan, and Yajing Shen. Soft magnetic skin for super-resolution tactile sensing with force self-decoupling. *Science Robotics*, 6(51):eabc8801, 2021.

Qi Ye, Qingtao Liu, Siyun Wang, Jiaying Chen, Yu Cui, Ke Jin, Huajin Chen, Xuan Cai, Gaofeng Li, and Jiming Chen. Visual-tactile pretraining and online multitask learning for humanlike manipulation dexterity. *Science Robotics*, 11(110):eady2869, 2026.

Ningbin Zhang, Jieji Ren, Yueshi Dong, Xinyu Yang, Rong Bian, Jinhao Li, Guoying Gu, and Xiangyang Zhu. Soft robotic hand with tactile palm-finger coordination. *Nature Communications*, 16(1):2395, 2025.

Zihang Zhao, Wanlin Li, Yuyang Li, Tengyu Liu, Boren Li, Meng Wang, Kai Du, Hangxin Liu, Yixin Zhu, Qining Wang, et al. Embedding high-resolution touch across robotic hands enables adaptive human-like grasping. *Nature Machine Intelligence*, 7(6):889–900, 2025.

Yuhang Zheng, Songen Gu, Weize Li, Yupeng Zheng, Yujie Zang, Shuai Tian, Xiang Li, Ce Hao, Chen Gao, Si Liu, Haoran Li, Yilun Chen, Shuicheng Yan, and Wenchao Ding. Omnivta: Visuotactile world modeling for contact-rich robotic manipulation, 2026. URL https://arxiv.org/abs/2603.19201.

A APPENDIX

CONTENTS

Figure S1: Embodied platform 14
A.1 Image feature extraction 14
A.2 MTNet optimization objective details 15
A.2.1 Reconstruction terms 15
A.2.2 Distributional alignment 15
A.2.3 Semantic alignment 16
A.2.4 Geometric alignment 17
A.3 AMTNet optimization objective details 17
Figure S2: Illustration of datasets 17
A.4 Dataset descriptions 18
A.5 Implementation details 18
Figure S3: Illustration of evaluation metrics 18
A.6 Evaluation metrics 19
A.6.1 Normalized root mean square error (NRMSE) 19
A.6.2 Spatial cosine similarity (S-CosSim) 19
A.6.3 Smoothed concordance correlation coefficient (S-CCC) 19
A.6.4 Temporal intersection over union (T-IoU) 20
A.6.5 Window-based F1 score (W-F1) 20
Figure S4: Loss curves of MTNet 20
Figure S5: Loss curves of tactile diffusion baseline 20
Figure S6: Cosine similarity of latent representations 20
Figure S7: AMTNet predictions on human and robotic images 20
Figure S8: Robotic mirror touch validation 20

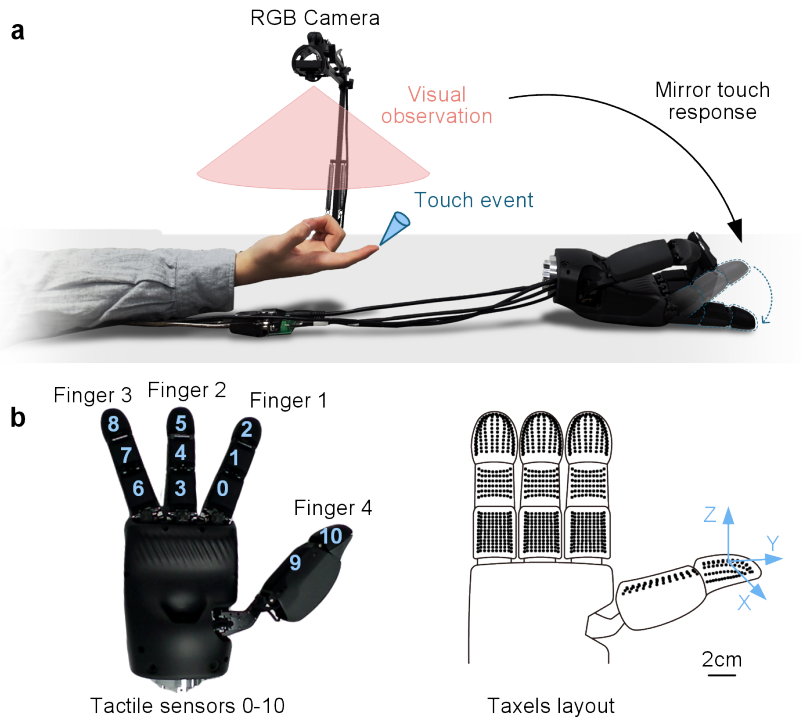


Fig. S1: **Embodied platform. a,** Our platform uses an RGB camera to capture human hand touch events, triggering corresponding tactile responses and reflexive behaviors in the PaXini DexH13 robotic hand. **b,** The robotic hand integrates 11 electromagnetic tactile sensors with a total of 1,140 taxels. Sensors 1, 4, and 7 each contain 60 taxels, while the remaining eight sensors each contain 120 taxels. The sensors provide 1-mm spatial resolution and three-axis force sensing.

A.1 Image feature extraction

This section details the feature extraction process for image frames in the MTNet visual pathway (Sec. 3.2). Given the visual input stack $I_r \in \mathbb{R}^{3K\times 224\times 224}$, each frame is first independently processed by a modified ResNet-18 backbone (He et al., 2016) to extract image features. This backbone discards the final downsampling and fully connected layers, projecting each frame into $S = 14 \times 14$ spatial patches. A linear layer then projects these patch features to a uniform latent dimension $d = 128$, yielding the spatial feature tensor $F \in \mathbb{R}^{K\times S\times d}$.

To aggregate motion cues across K frames, we apply a lightweight temporal attention fusion module. For each spatial patch $s \in \{1, ..., S\}$, a temporal score $a_{k,s}$ is computed across K frames, and the features are fused via a weighted sum:

$$a_{k,s} = \frac{\exp(F_{k,s}W_{\text{temp}})}{\sum_{j=1}^{K}\exp(F_{j,s}W_{\text{temp}})}, \quad \tilde{F} = \sum_{k=1}^{K} a_{k,s} \odot F_{k,s} \tag{10}$$

, where $W_{\text{temp}} \in \mathbb{R}^{d\times 1}$ is a learnable projection matrix and $\odot$ denotes element-wise multiplication. Note that the softmax is applied over the frame dimension for each patch independently. The resulting fused map $\tilde{F} \in \mathbb{R}^{S\times d}$ is augmented with learnable patch positional embeddings $E_{\text{pos}} \in \mathbb{R}^{S\times d}$ to form the visual memory $V = \tilde{F} + E_{\text{pos}}$.

Then, we define learnable spatial queries $Q \in \mathbb{R}^{P\times d}$ to assign patch features to $P = 11$ tactile areas. Multi-head cross-attention (MHCA) (Vaswani et al., 2017) collects relevant visual information from the visual memory V into each area:

$$H = \text{Softmax}\left(\frac{(QW_Q)(VW_K)^{\mathsf{T}}}{\sqrt{d}}\right)(VW_V) \tag{11}$$

, where $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$ are learnable projection matrices for queries, keys, and values, and $H \in \mathbb{R}^{P\times d}$ denotes the attended visual features used to compute Z_v in Eq. 1.

A.2 MTNET OPTIMIZATION OBJECTIVE DETAILS

This section specifies MTNet's 11 loss items introduced in Sec. 3.3: reconstruction terms ($\mathcal{L}^v_{\text{rec}}, \mathcal{L}^t_{\text{rec}}, \mathcal{L}^v_{\text{mag}}, \mathcal{L}^t_{\text{mag}}$), contact term ($\mathcal{L}_{\text{ct}}$), distributional alignment ($\mathcal{L}^v_{\text{KL}}, \mathcal{L}^t_{\text{KL}}, \mathcal{L}^{\mathcal{Q}_v\mathcal{Q}_t}_{\text{KL}}, \mathcal{L}^{\mathcal{P}_v\mathcal{Q}_t}_{\text{KL}}$), semantic alignment ($\mathcal{L}_{\text{con}}$), and geometric alignment ($\mathcal{L}_{\text{rel}}$).

A.2.1 RECONSTRUCTION TERMS

Let $O \in \mathbb{R}^{P\times N\times 3}$ denote the ground-truth force field, where P is the number of tactile regions, and N is the number of taxels per region. Let $M \in \{0,1\}^{P\times N}$ be the valid-taxel mask. The decoder produces two force fields, $\hat{O}_v$ and $\hat{O}_t$, as defined in Eq. 5. Here, the subscript $m \in \{v,t\}$ indexes the visual prediction ($m = v$) and tactile reconstruction ($m = t$) branches.

To emphasize taxels with physical contact, we define the taxel-wise contact indicator and the corresponding contact-aware weight map as:

$$\Delta = \mathbb{I}\left(\|O\|_2 > 0\right), \qquad \Omega = 1 + (\omega - 1)\Delta \tag{12}$$

, where the $\|\cdot\|_2$ norm is taken over the three force components and $\mathbb{I}(\cdot)$ denotes the indicator function. $\Delta \in \{0,1\}^{P\times N}$ marks taxels with nonzero ground-truth force, and $\Omega \in \mathbb{R}^{P\times N}$ assigns weight ω to contact taxels and weight 1 to noncontact taxels, with $\omega = 20$ in all experiments.

Thus, the weighted masked mean squared error (W-MSE) and the weighted masked mean absolute error (W-MAE) shown in Fig. 2c are defined as:

$$\mathcal{L}^m_{\text{rec}} = \frac{\sum M\Omega \left\|\hat{O}_m - O\right\|^2_2}{3\sum M\Omega + \epsilon}, \quad \mathcal{L}^m_{\text{mag}} = \frac{\sum M\Omega \left|\|\hat{O}_m\|_2 - \|O\|_2\right|}{\sum M\Omega + \epsilon}, \quad m \in \{v,t\} \tag{13}$$

, where ϵ is a small constant for numerical stability. The two sets of losses, $(\mathcal{L}^v_{\text{rec}}, \mathcal{L}^v_{\text{mag}})$ and $(\mathcal{L}^t_{\text{rec}}, \mathcal{L}^t_{\text{mag}})$, supervise the visual-to-tactile mapping and stabilizes the shared decoder through tactile self-reconstruction, respectively. $\mathcal{L}^m_{\text{mag}}$ enforces magnitude consistency by directly penalizing errors in contact magnitude, complementing the component-wise reconstruction objective of $\mathcal{L}^m_{\text{rec}}$.

The contact prediction loss shown in Fig. 2d uses region-level binary labels derived from the ground-truth force field. For region p, the binary contact label is defined as:

$$Y_{\text{ct},p} = \mathbb{I}\left(\sum_{n=1}^{N} M_{pn}\|O_{pn}\|_2 > \tau_{\text{ct}}\right) \tag{14}$$

, where τ_{ct} is the contact threshold. The predicted contact probabilities $\hat{c}$ from Eq. 1 are optimized with a class-weighted binary cross-entropy (BCE) loss:

$$\mathcal{L}_{\text{ct}} = -\frac{1}{P}\sum_{p=1}^{P}\left[\rho_{\text{ct}} Y_{\text{ct},p}\log(\hat{c}_p + \epsilon) + (1 - Y_{\text{ct},p})\log(1 - \hat{c}_p + \epsilon)\right] \tag{15}$$

, where ρ_{ct} compensates for the imbalance between contact and noncontact regions. This loss encourages visual pathway to encode contact location and supervises the contact modulation of $\hat{O}_v$.

A.2.2 DISTRIBUTIONAL ALIGNMENT

As defined in Eq. 3, the decoder generates distributions $(\mathcal{P}_v, \mathcal{P}_t, \mathcal{Q}_v, \mathcal{Q}_t)$ for both the visual pathway and the tactile pathway. Following the standard CVAE formulation (Sohn et al., 2015), we first minimize the KL divergence from the posterior to the prior within each branch:

$$\mathcal{L}^v_{\text{KL}} = D_{\text{KL}}\left(\mathcal{Q}_v\|\mathcal{P}_v\right), \qquad \mathcal{L}^t_{\text{KL}} = D_{\text{KL}}\left(\mathcal{Q}_t\|\mathcal{P}_t\right). \tag{16}$$

Beyond within-pathway CVAE regularization, MTNet leverages the tactile posterior as a probabilistic anchor for cross-modal distribution alignment (Fig. 2e). Since $\mathcal{Q}_t$ is inferred from observed

tactile measurements, it provides a more direct estimate of the underlying contact-force distribution than the visual branch alone. To reduce pathway-level noise, we maintain an exponential moving average (EMA) (Tarvainen & Valpola, 2017) of the tactile posterior parameters, yielding a smoothed anchor distribution $\bar{\mathcal{Q}}_t$. We then impose two cross-modal KL alignment losses:

$$\mathcal{L}_{\mathrm{KL}}^{Q_v Q_t} = D_{\mathrm{KL}}\left(\mathcal{Q}_v \| \bar{\mathcal{Q}}_t\right), \qquad \mathcal{L}_{\mathrm{KL}}^{P_v Q_t} = D_{\mathrm{KL}}\left(\mathcal{P}_v \| \bar{\mathcal{Q}}_t\right) \tag{17}$$

, where the first term aligns the visual posterior with the tactile posterior anchor during training when tactile data is available. The second term aligns the visual prior with the same anchor, an operation essential for inference because the stochastic latent variable is sampled from the visual prior when tactile observations are absent. Together, these losses transfer tactile distribution into the visual generative pathway.

The KL-related loss weights in Eq. 7 are introduced gradually with a linear ramp schedule to avoid destabilizing early reconstruction learning. For a target weight $\lambda_{\max}$, start epoch e_s, and end epoch e_e, the scheduled coefficient at epoch e is

$$\lambda(e) = \begin{cases} 0, & e < e_s, \\ \lambda_{\max} \dfrac{e - e_s}{e_e - e_s}, & e_s \le e < e_e, \\ \lambda_{\max}, & e \ge e_e. \end{cases} \tag{18}$$

This schedule allows the model to first learn stable force reconstruction and then progressively regularize the latent distributions.

A.2.3 Semantic alignment

MTNet aligns the deterministic visual tokens Z_t and tactile tokens Z_t (Fig. 2f), using a force-aware contrastive objective (Oord et al., 2018). Conventional contrastive learning treats only two views of the same sample as a positive pair. Our objective regards samples with similar measured force fields as positives, allowing their cross-modal tokens to share physically consistent semantics.

Let $\mathcal{I}$ denote the set of samples used to construct the contrastive pairs. For each sample $i \in \mathcal{I}$, we vectorize corresponding ground-truth force field O_i and normalize the resulting vector as $y_i = \frac{\mathrm{vec}(O_i)}{\|\mathrm{vec}(O_i)\|_2 + \epsilon}$. Thus, the force similarity between samples i and j is $A_{Y,ij} = y_i^{\mathsf{T}} y_j$. We identify force-similar pairs using a threshold τ_Y. The binary relation and the row-normalized positive distribution are defined as:

$$R_{Y,ij} = \mathbb{I}(A_{Y,ij} > \tau_Y) \vee \mathbb{I}(i = j), \quad \Pi_{Y,ij} = \frac{R_{Y,ij}}{\sum_{k \in \mathcal{I}} R_{Y,ik} + \epsilon} \tag{19}$$

, where the second term in $R_{Y,ij}$ ensures that each sample remains positive with itself, and $\Pi_{Y,ij}$ assigns equal probability to all force-similar positives. This distribution allows samples with comparable contact patterns to contribute to semantic alignment. The positive and negative pair sets are $\mathcal{P}_Y = \left\{(i,j) \in \mathcal{I}^2 : R_{Y,ij} = 1\right\}$ and $\mathcal{N}_Y = \left\{(i,j) \in \mathcal{I}^2 : R_{Y,ij} = 0\right\}$, respectively.

For each region p and modality pair $(a,b) \in \{(v,t),(t,v)\}$, the region-wise similarity logit between samples i and j is defined as:

$$s^p_{a \to b,ij} = \frac{\left(\bar{Z}^p_{a,i}\right)^{\mathsf{T}} \bar{Z}^p_{b,j}}{\tau_{\mathrm{con}}} \tag{20}$$

, where $\bar{Z}^p_{v,i} = \frac{Z^p_{v,i}}{\|Z^p_{v,i}\|_2 + \epsilon}$ and $\bar{Z}^p_{t,i} = \frac{Z^p_{t,i}}{\|Z^p_{t,i}\|_2 + \epsilon}$ are the normalized deterministic tokens, and τ_{con} is the contrastive temperature. Comparisons are restricted to tokens from the same tactile region, which preserves spatial correspondence between the two modalities.

Hence, the force-aware InfoNCE loss for one alignment direction is defined as:

$$\ell_{\mathrm{NCE}}(Z^p_a, Z^p_b; \mathcal{P}_Y, \mathcal{N}_Y, \Pi_Y) = -\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{I}} \Pi_{Y,ij} \log \frac{\exp\left(s^p_{a \to b,ij}\right)}{\sum_{k:\,(i,k) \in \mathcal{P}_Y \cup \mathcal{N}_Y} \exp\left(s^p_{a \to b,ik}\right)}. \tag{21}$$

By averaging this loss over both cross-modal directions and all tactile regions, we obtain the final semantic alignment loss $\mathcal{L}_{\mathrm{con}}$ in Eq. 6.

A.2.4 GEOMETRIC ALIGNMENT

We also align the relational geometry between visual and tactile tokens so that the two modalities preserve similar sample-to-sample topology in the shared latent space. The region-level tokens are first vectorized and normalized to form sample-level descriptors:

$$R_v = \frac{\mathrm{vec}(Z_v)}{\|\mathrm{vec}(Z_v)\|_2 + \epsilon}, \qquad R_t = \frac{\mathrm{vec}(Z_t)}{\|\mathrm{vec}(Z_t)\|_2 + \epsilon}. \tag{22}$$

The relational geometry loss is then defined as:

$$\mathcal{L}_{\mathrm{rel}} = \frac{1}{B^2} \left\| R_v R_v^{\mathrm{T}} - R_t R_t^{\mathrm{T}} \right\|_F^2, \tag{23}$$

where $\| \cdot \|_F$ denotes the Frobenius norm and B is the batch size.

A.3 AMTNET OPTIMIZATION OBJECTIVE DETAILS

This section formulates the four cross-domain alignment losses $\mathcal{L}^{\mathrm{A}}_{\mathrm{feat}}$, $\mathcal{L}^{\mathrm{A}}_{\mathrm{KL}}$, $\mathcal{L}^{\mathrm{A}}_{\mathrm{ct}}$, and $\mathcal{L}^{\mathrm{A}}_{\mathrm{pred}}$ in AMTNet (Sec. 3.4). As shown in Fig. 3, AMTNet is trained on paired image sequences (I_h, I_r) with matched contact configurations. The robot output serves as the teacher, while the human output serves as the student. Both outputs are obtained as:

$$\left(Z_r, \mathcal{P}_r, \hat{c}_r, \hat{O}_r\right) = \mathrm{AMTNet}(I_r), \quad \left(Z_h, \mathcal{P}_h, \hat{c}_h, \hat{O}_h\right) = \mathrm{AMTNet}(I_h) \tag{24}$$

, while Z_r and Z_h denote the deterministic visual tokens, $\mathcal{P}_r$ and $\mathcal{P}_h$ are the visual priors, $\hat{c}_r$ and $\hat{c}_h$ contain the predicted contact probabilities, $\hat{O}_r$ and $\hat{O}_h$ are the predicted force fields, respectively.

Thus, each objective term aligns a student output from the human stream with the corresponding detached teacher output from the robot stream, using the stop-gradient operator $\eta(\cdot)$:

$$\begin{aligned} \mathcal{L}^{\mathrm{A}}_{\mathrm{feat}} &= \|Z_h - \eta(Z_r)\|_2^2, \quad & \mathcal{L}^{\mathrm{A}}_{\mathrm{KL}} &= D_{\mathrm{KL}}\big(\mathcal{P}_h \,\|\, \eta(\mathcal{P}_r)\big) \\ \mathcal{L}^{\mathrm{A}}_{\mathrm{ct}} &= \mathrm{BCE}(\hat{c}_h, \eta(\hat{c}_r)), \quad & \mathcal{L}^{\mathrm{A}}_{\mathrm{pred}} &= \big\|\hat{O}_h - \eta(\hat{O}_r)\big\|_2^2, \end{aligned} \tag{25}$$

where the feature alignment loss $\mathcal{L}^{\mathrm{A}}_{\mathrm{feat}}$ transfers the semantic structure learned by MTNet to the human visual encoder, ensuring that human hand images are mapped into the same tactile-aware latent space as robot hand images. The KL alignment loss $\mathcal{L}^{\mathrm{A}}_{\mathrm{KL}}$ aligns the human and robot priors, thereby transferring the stochastic generative capabilities learned from robot data to the human domain for tactile inference. The contact alignment loss $\mathcal{L}^{\mathrm{A}}_{\mathrm{ct}}$ and prediction alignment loss $\mathcal{L}^{\mathrm{A}}_{\mathrm{pred}}$ preserve the generative behavior of the frozen decoder by enforcing consistency between the human and robot tactile predictions. In practice, we use the contact logit rather than a hard binary label for $\mathcal{L}^{\mathrm{A}}_{\mathrm{ct}}$, which provides smoother supervision. The four alignment losses above, together with the gate loss in Eq. 8, constitute the total optimization objective of AMTNet in Eq. 9.

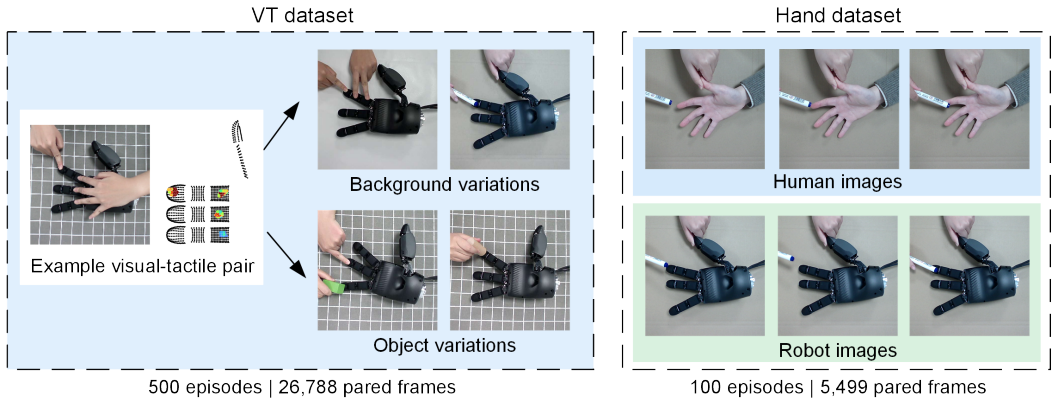


Fig. S2: **Illustration of datasets.** The VT and Hand datasets are used to train and evaluate MTNet and AMTNet, respectively.

A.4 Dataset descriptions

Two datasets were collected for model training and evaluation (Fig. S2): the VT dataset for MTNet and the Hand dataset for AMTNet. Experiments in Sec. 4.1 are performed on the VT dataset, and those in Sec. 4.3 on the Hand dataset. The VT dataset was collected on the robotic hand by performing diverse touch interactions with synchronous recording visual frames and tactile signals. Each visual frame measures $224 \times 224 \times 3$, and each tactile frame measures 1140×3. The VT dataset consists of 500 episodes totaling 26,788 aligned visual–tactile frame pairs. The Hand dataset was created by selecting robotic touch images from the VT dataset and manually reproducing the same touch poses with a human hand for image capture. This resulted in a paired visual dataset of robot and human hands, containing 100 episodes comprising 5,499 image pairs. Both datasets are split based on episodes into training and validation sets at an 8:2 ratio.

A.5 Implementation details

MTNet was trained on VT dataset for 200 epochs using $K = 5$ consecutive RGB frames as visual input. We employed the AdamW optimizer (Loshchilov & Hutter, 2017) with an initial learning rate of 4×10^{-4} and a weight decay of 10^{-5}. A cosine annealing learning rate scheduler was applied with a minimum learning rate of 10^{-6}. For the force-aware contrastive loss, the temperature and force similarity threshold were set to 0.07 and 0.85, respectively. Subsequently, AMTNet was initialized from the pretrained MTNet checkpoint and trained on the Hand dataset for 500 epochs. Only the human visual encoder and the gating network were optimized, while the pretrained components were kept frozen. This stage utilized $K = 5$ frames as input and the AdamW optimizer with a learning rate of 10^{-3}, following a cosine annealing scheduler with a 10^{-7} minimum. All models were trained and evaluated on four NVIDIA 80GB A100 GPUs.

Tactile diffusion baseline. We implement a tactile diffusion baseline (Zheng et al., 2026), which predicts tactile responses from visual observations through a strong denoising objective, without explicitly enforcing cross-modal alignment between visual and tactile representations. The original model predicts image-like tactile signals characteristic of vision-based tactile sensors. To adapt it to our independent taxel-based setting, we make two modifications: **(i)** we replace the tactile image with a 1,140-dimensional taxel representation and condition the prediction on the 3D coordinates of each taxel; **(ii)** we restrict the prediction target to the tactile response at the K-th time step, conditioned on the same K consecutive visual frames as used in our method. All other model components, optimization objectives, and the two-stage training strategy follow the original work. We pretrain the TactileVAE for 100 epochs and train the diffusion model for 300 epochs, respectively. Parameters not specified in the original work are set conservatively, and we adopt the same preprocessing and post-processing procedures as in our method to ensure a fair comparison.

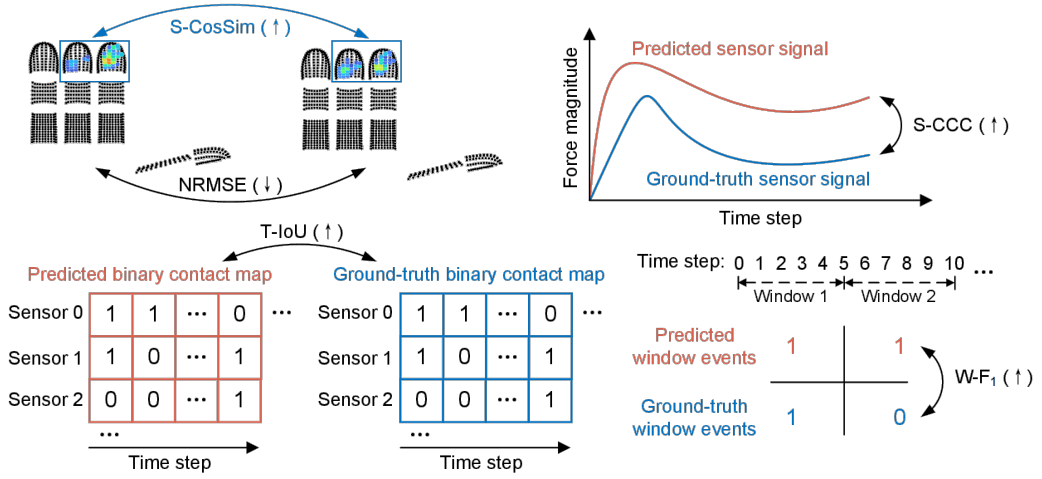


Fig. S3: **Illustration of evaluation metrics.** These metrics are used to assess tactile prediction performance (Sec. 4.1), with computation details in Appendix A.6.

A.6 EVALUATION METRICS

This section details the five metrics used to quantitatively evaluate the prediction performance of MTNet (Sec. 4.1). We evaluate our model from both spatial and temporal perspectives (Fig. S3), covering taxel-wise force reconstruction error, spatial force-pattern agreement, temporal force consistency, contact-state accuracy, and touch event detection. These metrics are designed to address **(i)** the sparsity of tactile events, which can bias standard evaluations, and **(ii)** frame jitter, which can lead to temporal inconsistency. To avoid sensor noise, we apply a noise floor threshold to both predicted and ground-truth force fields before computing these metrics. All metrics are computed per sample and then averaged to produce the final evaluation.

A.6.1 NORMALIZED ROOT MEAN SQUARE ERROR (NRMSE)

We compute NRMSE to quantify the taxel-wise force prediction error relative to the force range, obtained by first computing the root mean square error and then normalizing by the ground-truth force range:

$$\mathrm{RMSE} = \sqrt{\frac{1}{3 \times 1140 \times T} \sum_{t=1}^{T} \left\| O_t - \hat{O}_t \right\|_2^2}, \quad \mathrm{NRMSE} = \frac{\mathrm{RMSE}}{\max_t O_t - \min_t O_t} \tag{26}$$

, where T is the sequence length; $\hat{O}_t$ and O_t denote the predicted and ground-truth force fields at time step t, respectively. In cases where the ground-truth force range is zero, NRMSE is defined as 0 if the prediction is also zero, and 1 otherwise. This metric directly evaluates the numerical accuracy of the predicted 3D tactile force vectors.

A.6.2 SPATIAL COSINE SIMILARITY (S-COSSIM)

Spatial cosine similarity (S-CosSim) is employed to assess whether the predicted force field preserves the spatial pattern of the ground-truth. To reduce sensitivity to one-frame temporal offsets, each ground-truth frame is matched to the most similar predicted frame within a local temporal neighborhood:

$$s_t = \max_{\tau \in [t-\Delta, t+\Delta]} \frac{\mathrm{vec}(O_t)^T \, \mathrm{vec}(\hat{O}_t)_\tau}{\| \mathrm{vec}(O_t) \|_2 \| \mathrm{vec}(\hat{O}_t)_\tau \|_2 + \epsilon}, \qquad \Delta = 1 \tag{27}$$

, where $s_t \in [-1, 1]$ is the best local cosine similarity for ground-truth frame t, τ indexes candidate frames within the temporal window, and $\Delta = 1$ allows one frame tolerance. Then, S-CosSim is computed using force-energy weights:

$$\text{S-CosSim} = \frac{\sum_t \omega_t s_t}{\sum_t \omega_t}, \quad \omega_t = 0.1 + 0.9 \frac{\| \mathrm{vec}(O_t) \|_2}{\max_t \| \mathrm{vec}(O_t) \|_2 + \epsilon} \tag{28}$$

, where $\omega_t \in \mathbb{R}$ is a force-energy weight that assigns greater importance to frames with stronger force responses. S-CosSim complements NRMSE by focusing on the geometric consistency of the predicted force distribution rather than solely on its numerical error.

A.6.3 SMOOTHED CONCORDANCE CORRELATION COEFFICIENT (S-CCC)

We evaluate the temporal consistency of the predicted force using a smoothed concordance correlation coefficient (S-CCC), a smoothed version of the standard concordance correlation coefficient. For each of the 11 tactile regions, this metric assesses whether the resultant force trend over time matches the ground-truth sequence. Let $\hat{F}_t^p \in \mathbb{R}$ and $F_t^p \in \mathbb{R}$ denote the predicted and ground-truth resultant force magnitudes for region p at time step t, computed from $\hat{O}_t$ and O_t, respectively. To improve robustness against temporal jitter and emphasize contact-active intervals, we first smooth both sequences using a one-dimensional Gaussian filter and then compute a force-dependent temporal weight for each frame:

$$w_t^p = 0.1 + 0.9 \frac{\tilde{F}_t^p}{\max \tilde{F}_t^p + \epsilon}, \quad \tilde{F}^p = \mathcal{G}_\sigma \left(F^p \right), \quad \tilde{\hat{F}}^p = \mathcal{G}_\sigma \left(\hat{F}^p \right) \tag{29}$$

, where $\mathcal{G}_\sigma(\cdot)$ denotes Gaussian smoothing along the temporal dimension with $\sigma = 2$, and $w_t^p \in \mathbb{R}$ is the temporal weight for region p at time step t. Hence, the weighted temporal mean of smoothed sequences can be computed as:

$$\bar{F}^p = \frac{\sum_t w_t^p \tilde{F}_t^p}{\sum_t w_t^p}, \quad \bar{\hat{F}}^p = \frac{\sum_t w_t^p \tilde{\hat{F}}_t^p}{\sum_t w_t^p} \tag{30}$$

, where $\bar{F}^p$ and $\bar{\hat{F}}^p$ are the weighted means of the smoothed sequences $\tilde{F}^p$ and $\tilde{\hat{F}}^p$, respectively. The final S-CCC score is obtained by averaging the weighted concordance correlation across all regions:

$$\text{S-CCC} = \frac{1}{P}\sum_{p=1}^{P} \frac{2\sum_t w_t^p (\tilde{F}_t^p - \bar{F}^p)(\tilde{\hat{F}}_t^p - \bar{\hat{F}}^p)}{\sum_t w_t^p \left[(\tilde{F}_t^p - \bar{F}^p)^2 + (\tilde{\hat{F}}_t^p - \bar{\hat{F}}^p)^2\right] + \sum_t w_t^p (\bar{F}^p - \bar{\hat{F}}^p)^2 + \epsilon}. \tag{31}$$

If both ground-truth and predicted sequences exhibit zero force, the regional score is defined as 1. S-CCC quantifies the concordance of temporal trends and amplitudes, emphasizing physically meaningful contact periods to reflect prediction performance.

A.6.4 Temporal intersection over union (T-IoU)

Temporal intersection over union (T-IoU) evaluates contact-state overlap based on region-level binary contact labels. Following the definitions of F_t^p and $\hat{F}_t^p$ in S-CCC, the contact labels are obtained by thresholding the region-level force magnitudes:

$$y_t^p = \mathbb{I}\left(F_t^p > \tau_{\text{ct}}\right), \qquad \hat{y}_t^p = \mathbb{I}\left(\hat{F}_t^p > \tau_{\text{ct}}\right) \tag{32}$$

, where $y_t^p, \hat{y}_t^p \in \{0, 1\}$ denote the ground-truth and predicted binary contact states for region p at time step t, $\mathbb{I}(\cdot)$ is the indicator function, and τ_{ct} is the contact threshold. T-IoU is then computed over all time steps and tactile regions:

$$\text{T-IoU} = \frac{\sum_{t,p} \mathbb{I}\left(y_t^p = 1 \wedge \hat{y}_t^p = 1\right)}{\sum_{t,p} \mathbb{I}\left(y_t^p = 1 \vee \hat{y}_t^p = 1\right)}. \tag{33}$$

If both ground-truth and predicted sequences contain no contact, T-IoU is set to 1. This metric quantifies whether the predicted contact activations occur in the correct tactile regions and at the correct time steps.

A.6.5 Window-based F1 score (W-F1)

We use a window-based F1 score (W-F1) to further quantify the temporal accuracy of contact events. The tactile sequence is partitioned into non-overlapping windows $\{\mathcal{W}_u\}_{u=1}^{U}$ of five frames each, with zero-padding applied to any final segment shorter than five frames. Windows accommodate minor timing offsets between predicted and ground-truth responses. Based on the contact labels from Eq. 32, the window-level labels for each region p are defined as:

$$Y_u^p = \max_{t \in \mathcal{W}_u} y_t^p, \qquad \hat{Y}_u^p = \max_{t \in \mathcal{W}_u} \hat{y}_t^p \tag{34}$$

, where $Y_u^p, \hat{Y}_u^p \in \{0, 1\}$ indicate whether at least one contact event occurs within window $\mathcal{W}_u$. Treating each window as a binary classification instance, we compute the per-region F1 score from the true positives TP^p, false positives FP^p, and false negatives FN^p aggregated across all windows. The overall W-F1 is then obtained by averaging over all P regions:

$$\text{W-F1} = \frac{1}{P}\sum_{p=1}^{P} \frac{2\text{TP}^p}{2\text{TP}^p + \text{FP}^p + \text{FN}^p}. \tag{35}$$

If a region has no positive contacts in both the ground truth and the prediction across all windows, its F1 score is set to 1. W-F1 measures event-level contact detection accuracy, complementing T-IoU by being more robust to minor temporal misalignment.

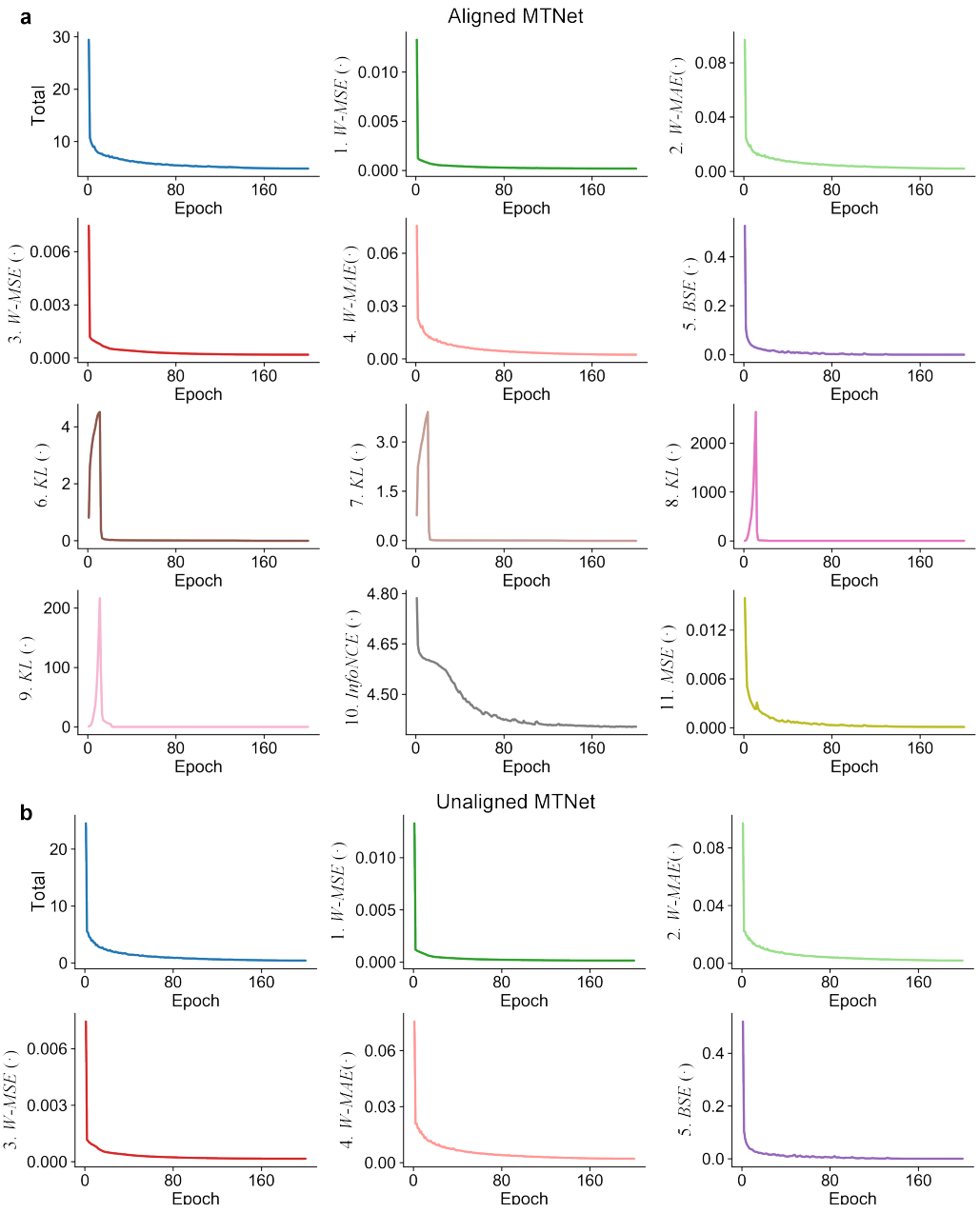


Fig. S4: **Loss curves of MTNet. a,** Eleven losses of the fully trained MTNet, ordered as in Fig. 2c-g. **b,** Losses of the unaligned MTNet trained with only the five reconstruction losses.

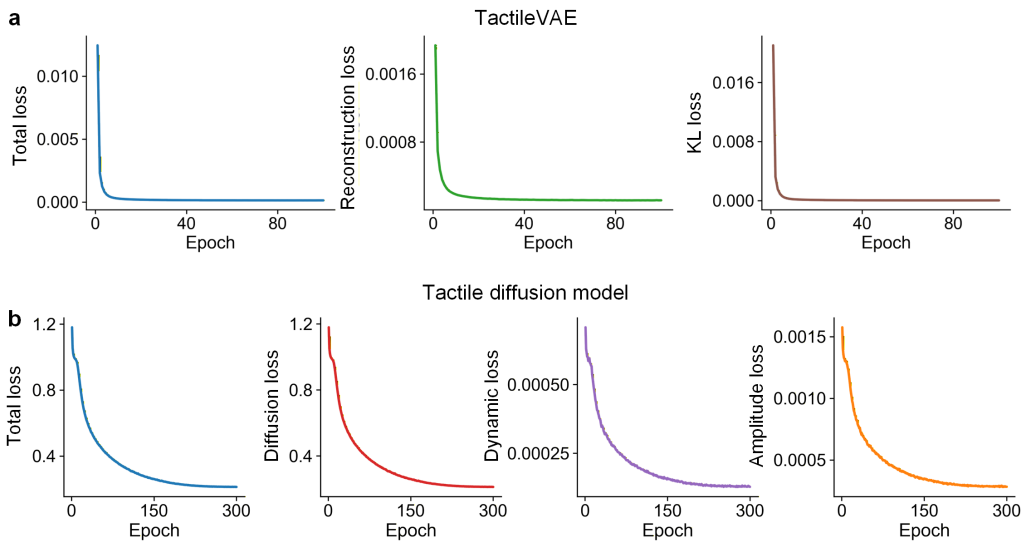


Fig. S5: **Loss curves of tactile diffusion baseline.** The baseline follows the original two-stage strategy: it first pretrains the TactileVAE and then trains the diffusion model, using the same loss terms as those defined in (Zheng et al., 2026). **a,** Loss curves for pretraining TactileVAE. **b,** Loss curves for training tactile diffusion model.

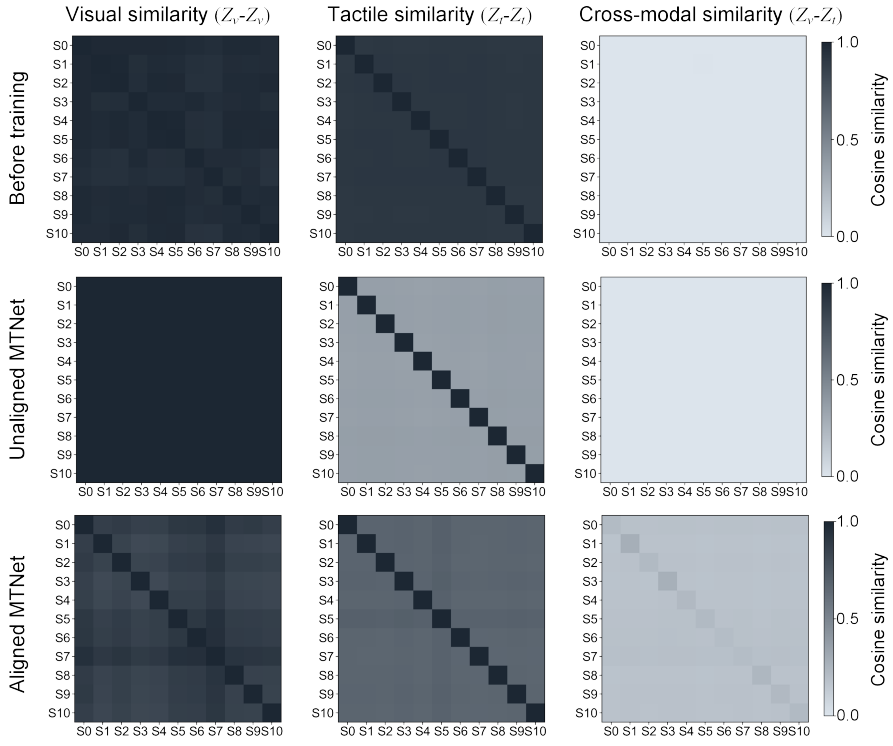


Fig. S6: **Cosine similarity of latent representations.** Matrices show within- and cross-modal similarities for visual (Z_v) and tactile (Z_t) tokens across the three MTNet variants in Fig. 5.

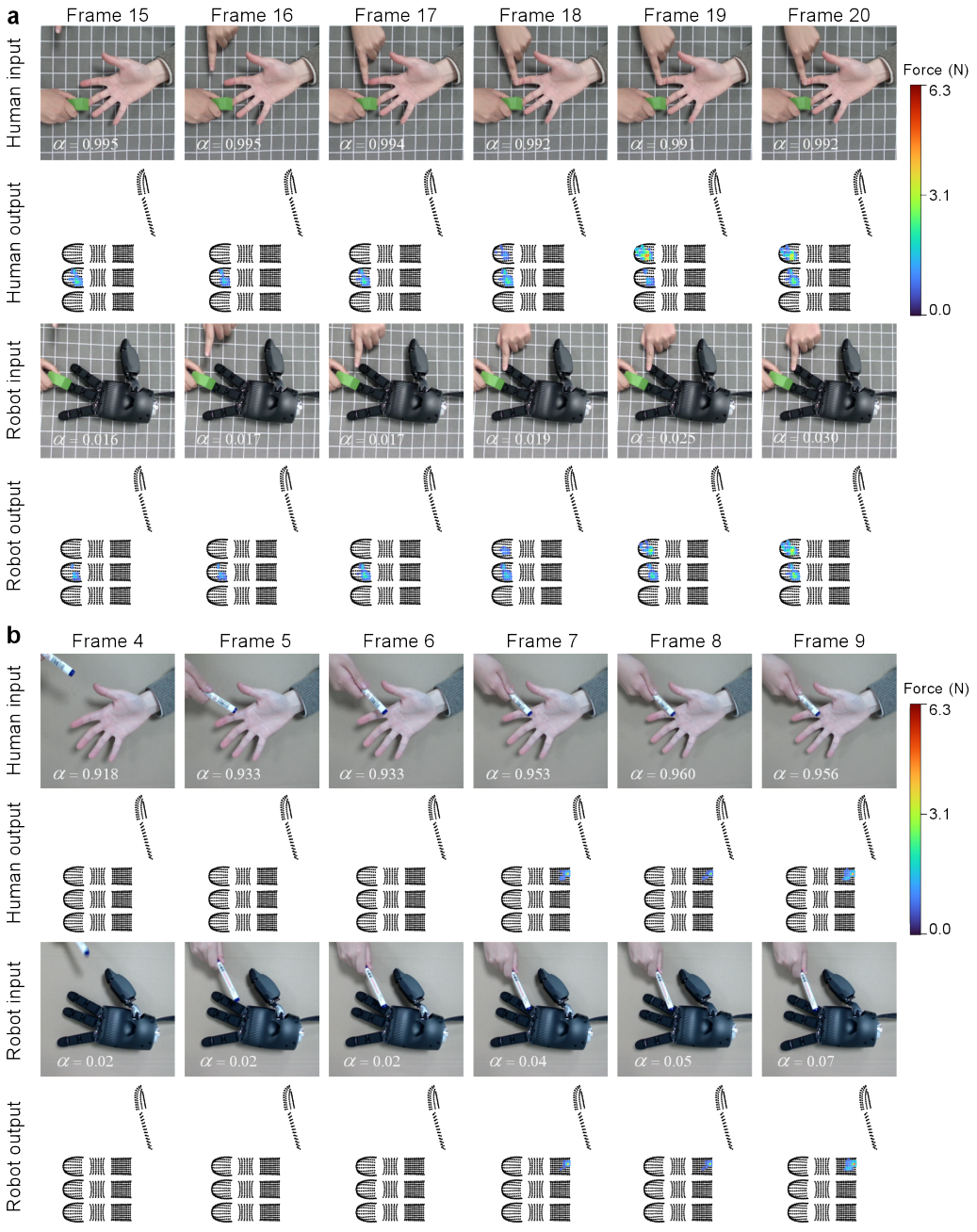


Fig. S7: **AMTNet predictions on human and robot images. a,** Predictions for both human and robotic images, expanded from Fig. 6c. **b,** Another sample. See Supplementary Video 2 for more demonstrations.

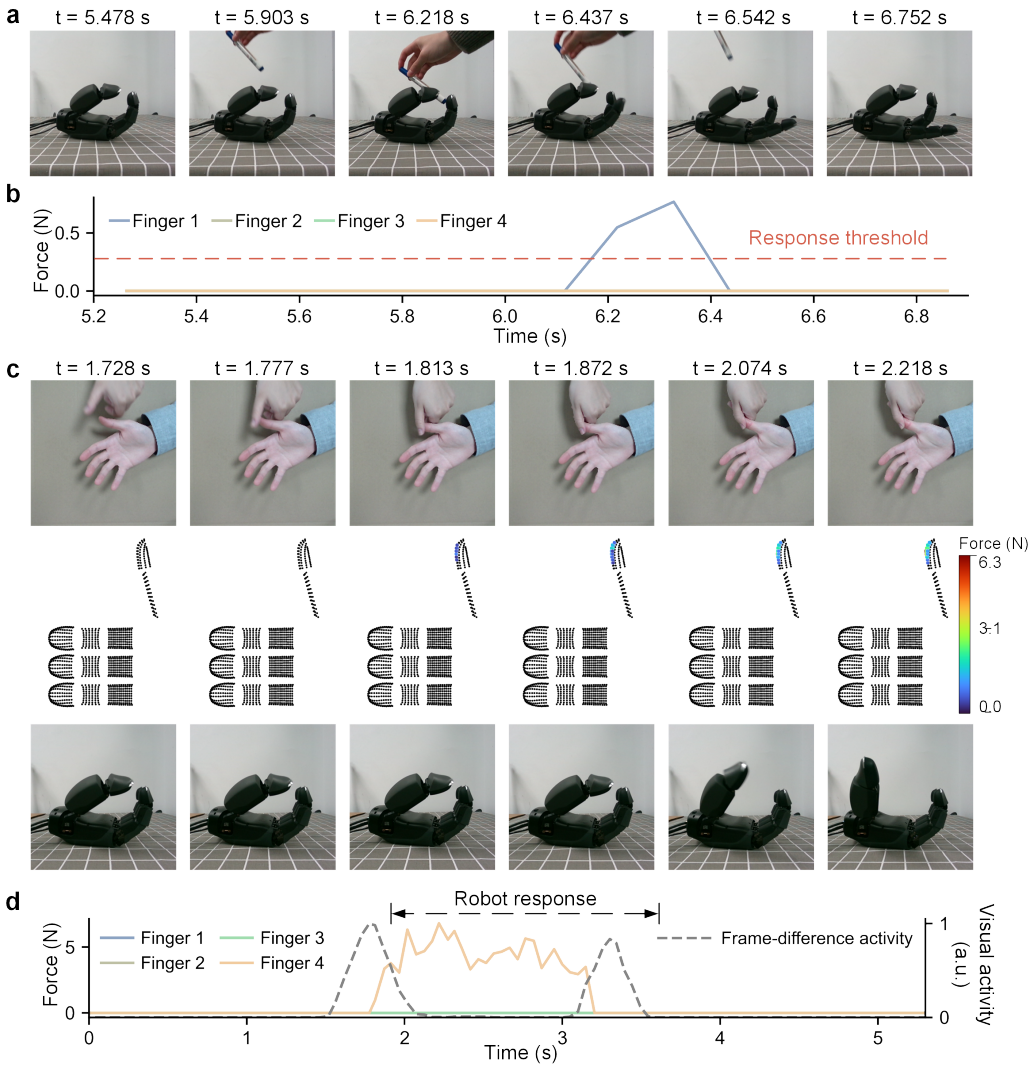


Fig. S8: **Robotic mirror touch validation.** **a,** Robotic tactile reflex setup, in which a predefined controller triggers a flick when the finger's maximum force exceeds 0.2 N and reset after 0.5 s without tactile input. **b,** Maximum force magnitude traces for the four fingers in the setup of **a**. **c,** AMTNet predicts tactile signals from 30 Hz images of a touched human hand, eliciting finger responses in an untouched robotic hand. **d,** Predicted four-finger tactile signals in **c** overlaid with frame-difference motion curves, illustrating temporal correspondence with visual touch. See Supplementary Video 3 for unseen object and extended demonstrations.