HAND KINEMATICS AS AN IMPLICIT PHYSICAL SENSOR FOR EGOCENTRIC GRASP PRESSURE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating grasp pressure from egocentric video is a partially observed and underdetermined inverse contact problem, as deformation at the contact interface is often occluded. Existing methods rely mainly on RGB appearance and overlook hand motion as a source of physical evidence. We present KIPS, which treats hand kinematics as an implicit physical sensor. During stable grasp transport, hand and object motion are mechanically coupled, making the recovered hand trajectory informative about the contact load required to support and accelerate the object. KIPS represents motion through displacement, velocity, and acceleration at multiple granularities, then learns an implicit mapping from these observations to dense sensor responses. RGB features optionally condition the model with complementary appearance cues. Across EgoTactile and OpenTouch, KIPS without direct RGB appearance conditioning outperforms appearance-based baselines on most metrics, while adding RGB features yields further improvements. Motion-stratified analysis shows that stronger motion coincides with larger relative pressure changes and smaller gains from RGB, underscoring the value of kinematic evidence in dynamic interactions.

1 INTRODUCTION

Estimating grasp pressure from egocentric video requires recovering both where the hand contacts an object and how strongly it presses (Grady et al., 2021; Christen et al., 2022; Collins et al., 2023; Mandikal & Grauman, 2021; Meier et al., 2021). This information reveals how a person physically loads an object and is important for robot learning from human demonstrations, where visually similar grasps may produce different interaction outcomes (Banerjee et al., 2025; Chao et al., 2021). It is also valuable for mixed reality and teleoperation, where contact intensity can support haptic feedback and pressure sensitive interaction (Henderson & Feiner, 2010; Sundaram et al., 2019; Wang et al., 2025; Jeon et al., 2026).

Early vision-based studies measured pressure on instrumented planar surfaces. PressureVisionDB (Grady et al., 2022), ContactLabelDB (Grady et al., 2024), and EgoPressure (Zhao et al., 2025) capture actions such as pressing, tapping, and sliding on pressure sensing panels synchronized with cameras, providing dense and metric pressure measurements. However, planar interactions do not represent the distributed contact that occurs when a hand grasps a three dimensional object. More recent datasets, including EgoTactile (Zeng et al., 2026) and OpenTouch (Song et al., 2025), use sensorized gloves to record taxel pressure during everyday object manipulation. These datasets have enabled models that estimate contact and pressure from RGB video using discriminative appearance representations or generative reconstruction. Nevertheless, appearance alone provides limited physical evidence because deformation at the contact region is often occluded by the grasped object. Similar visible appearances may therefore correspond to different pressure distributions (Brahmbhatt et al., 2020; Corona et al., 2020).

We approach this partially observed inverse contact problem through hand kinematics. During stable grasp transport, the motions of the hand and object are mechanically coupled. Under a simplified rigid-object model, Newton's second law requires the hand to generate a resultant contact force satisfying $\mathbf{F}_{\text{contact}} + m\mathbf{g} = m\mathbf{a}$, linking contact loading to the observed motion (Flanagan & Wing, 1993; Gao et al., 2005). This coupling makes the recovered hand trajectory informative at multiple scales. Global wrist motion reflects hand-object transport, palm-relative finger motion provides cues

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.

about where contact is being established or released, and the continuous trajectory provides temporal context for the progression from approach through sustained contact to release. Hand kinematics therefore provides physically grounded evidence about the underlying pressure distribution. Internal grip forces may vary without altering the object motion. We thus treat dense pressure recovery as conditional inference rather than exact mechanics inversion.

Based on this perspective, we introduce KIPS, which treats hand kinematics as an implicit physical sensor for egocentric grasp pressure estimation. KIPS uses inter-frame motion as its primary input and represents displacement, velocity, and acceleration at multi-granularities. Wrist motion captures the transport shared by the hand and object, while joint motion relative to the palm captures local articulation. An acceleration prototype dictionary captures recurring motion patterns, while a phase-structured temporal module models pressure evolution from approach through contact to release. The resulting joint-level and temporal representations are decoded at fixed taxel or mesh locations to recover dense pressure. KIPS operates without direct RGB appearance conditioning, but can incorporate RGB features to resolve residual ambiguity.

Experiments on EgoTactile and OpenTouch show that kinematics alone outperforms appearance-based methods on most evaluation metrics, establishing motion as a strong primary modality for grasp pressure estimation. Incorporating appearance features yields further improvements, demonstrating the empirical complementarity of the two modalities. Motion-stratified analyses further show larger relative pressure changes and diminishing RGB gains on both datasets as motion strengthens, consistent with greater kinematic observability during dynamic interactions.

Our contributions are summarized as follows:

- We formulate egocentric grasp pressure estimation as a partially observed, underdetermined inverse contact problem, with hand kinematics providing physically grounded evidence for dense pressure recovery. To our knowledge, this work presents the first systematic study of hand kinematics as the primary signal for egocentric grasp pressure estimation.
- We propose KIPS, a kinematics-led framework that combines multi-granularity motion representation, phase-structured temporal modeling, and identity-guided pressure decoding. KIPS estimates dense pressure without direct RGB appearance conditioning and optionally incorporates RGB as complementary context.
- Extensive experiments on EgoTactile and OpenTouch show that KIPS using kinematics alone outperforms appearance-based baselines on most metrics, while incorporating RGB yields further improvements and state-of-the-art performance. Motion-stratified analyses further reveal a systematic association between motion intensity and pressure variation.

2 Related Work

2.1 Vision-based hand contact and pressure estimation

Vision-based methods have progressed from contact localization to dense pressure reconstruction. HACO (Jung & Lee, 2025) predicts dense contact over the hand mesh but does not recover pressure magnitude. PressureVision (Grady et al., 2022) regresses pressure on a planar sensing surface from a single RGB image. PressureFormer (Zhao et al., 2025) extends pressure estimation to egocentric views by using MANO (Romero et al., 2017) vertices to query image features and projecting the predictions onto a hand UV map. EgoTactile (Zeng et al., 2026) introduces full-hand pressure estimation for everyday object interactions, with EgoPressureFormer performing discriminative video regression and EgoPressureDiff using generative priors to complete pressure in occluded regions. OpenTouch (Song et al., 2025) studies a related cross-modal setting by aligning video, pose, and tactile sequences for retrieval rather than directly regressing per-frame pressure.

Despite these advances, existing dense pressure estimators remain predominantly appearance-driven (Brahmbhatt et al., 2019; Jeon et al., 2026). PressureFormer uses hand geometry primarily for spatial querying, whereas OpenTouch treats pose sequences as a retrieval modality. Neither explicitly models inter-frame kinematics as physically informative evidence for per-frame dense pressure estimation. KIPS instead represents global wrist transport and palm-relative joint dynamics to recover dense pressure. It can operate without direct RGB appearance conditioning while optionally using appearance to provide complementary visual context.

2.2 MOTION-BASED FORCE INFERENCE

Motion has been used to infer interaction forces through inverse dynamics and learned physical models. (Pham et al., 2018) estimate fingertip contact forces from markerless hand-object tracking by combining Newton-Euler dynamics with a learned model of internal grip forces. (Li et al., 2019) jointly optimize human and object motion, contact locations, and interaction forces under whole-body dynamics. PhysPT (Zhang et al., 2024) refines human motion and infers joint actuation and ground contact forces using physics-aware representations and Euler-Lagrange constraints. Related multimodal systems combine hand configuration with physiological sensing. PiMForce (Seo et al., 2024) fuses 3D hand posture with forearm sEMG for whole-hand pressure estimation, while AET-TacNet (Pan et al., 2026) combines kinematic measurements and EMG to infer tactile responses.

While effective in their intended settings, these methods do not directly address dense pressure estimation in everyday egocentric interactions (Ehsani et al., 2020; Rempe et al., 2020). Dynamics-based approaches rely on tracked object motion or parameterized body dynamics and recover net wrenches, sparse contact forces, or ground reactions rather than pressure over the hand surface. Physiological approaches can estimate pressure or tactile responses but require EMG and dedicated sensors. KIPS instead uses hand kinematics as the primary observation for dense taxel- or vertex-level pressure prediction, without known mass or inertia, or physiological sensing. It handles the remaining ambiguity through conditional inference rather than explicit dynamics inversion.

3 KIPS

KIPS maps recovered hand kinematics, optionally augmented with RGB appearance cues, to dense pressure sequences. Given an egocentric video, we recover MANO trajectories using HaWoR (Zhang et al., 2025) and derive global wrist motion and palm-relative joint dynamics. As illustrated in Figure 1, KIPS comprises three components: Inter-Frame Kinematic Representation (IKR), Phase-Structured Temporal Dynamics (PSTD), and Identity-Guided Pressure Decoding (IGPD). The model operates on overlapping temporal windows, using 64 frames with stride 32.

3.1 INTER-FRAME KINEMATIC REPRESENTATION

IKR encodes motion at joint and wrist granularities. The joint stream captures local articulation in the palm coordinate system, while the wrist stream describes the global transport and configuration.

Multi-granularity kinematic observations. Let $\mathbf{q}_{t,j} \in \mathbb{R}^3$ denote the position of MANO joint j at time t, recovered through forward kinematics, where $j = 1, \ldots, J$ and $J = 21$. We express each joint relative to the wrist coordinate system as

$$\mathbf{d}_{t,j} = \frac{R_t^\top (\mathbf{q}_{t,j} - \mathbf{q}_{t,w})}{\ell_t}, \qquad \mathbf{v}_{t,j} = \mathcal{D}_\tau \mathbf{d}_{t,j}, \qquad \mathbf{a}_{t,j} = \mathcal{D}_\tau \mathbf{v}_{t,j}, \tag{1}$$

where R_t is the wrist orientation, $\mathbf{q}_{t,w}$ is the wrist position, ℓ_t is a hand scale factor, and $\mathcal{D}_\tau$ denotes differentiation using the observed frame timestamps, using one-sided differences at the start and end of the recorded sequence. This normalization removes global translation and scale from local articulation. We further include frame-level changes in velocity and acceleration:

$$\Delta\mathbf{v}_{t,j} = \mathbf{v}_{t,j} - \mathbf{v}_{t-1,j}, \qquad \Delta\mathbf{a}_{t,j} = \mathbf{a}_{t,j} - \mathbf{a}_{t-1,j}. \tag{2}$$

The complete joint observation is given as follows:

$$\mathbf{x}_{t,j} = [\mathbf{d}_{t,j}, \mathbf{v}_{t,j}, \mathbf{a}_{t,j}, \Delta\mathbf{v}_{t,j}, \Delta\mathbf{a}_{t,j}] \in \mathbb{R}^{15}. \tag{3}$$

The wrist stream retains global motion and hand configuration, which is defined as

$$\mathbf{b}_t = [\mathbf{d}_{t,w}, \mathbf{v}_{t,w}, \mathbf{a}_{t,w}, \boldsymbol{\xi}_t] \in \mathbb{R}^{60}, \qquad \boldsymbol{\xi}_t = \left[R_t, \boldsymbol{\theta}_t, R_t^\top \mathbf{g}_{\text{world}}\right] \in \mathbb{R}^{51}, \tag{4}$$

where $\mathbf{d}_{t,w}$, $\mathbf{v}_{t,w}$, and $\mathbf{a}_{t,w}$ describe wrist displacement, velocity, and acceleration. The auxiliary state $\boldsymbol{\xi}_t$ contains the global wrist orientation R_t, MANO pose rotations θ_t, and gravity expressed in the wrist coordinate system $R_t^\top \mathbf{g}_{\text{world}}$. Joint-level kinematics capture local articulation, whereas wrist-level kinematics describe the global hand motion during grasp transport. Separate MLP encoders project the two streams into a 128 dimensional latent space.

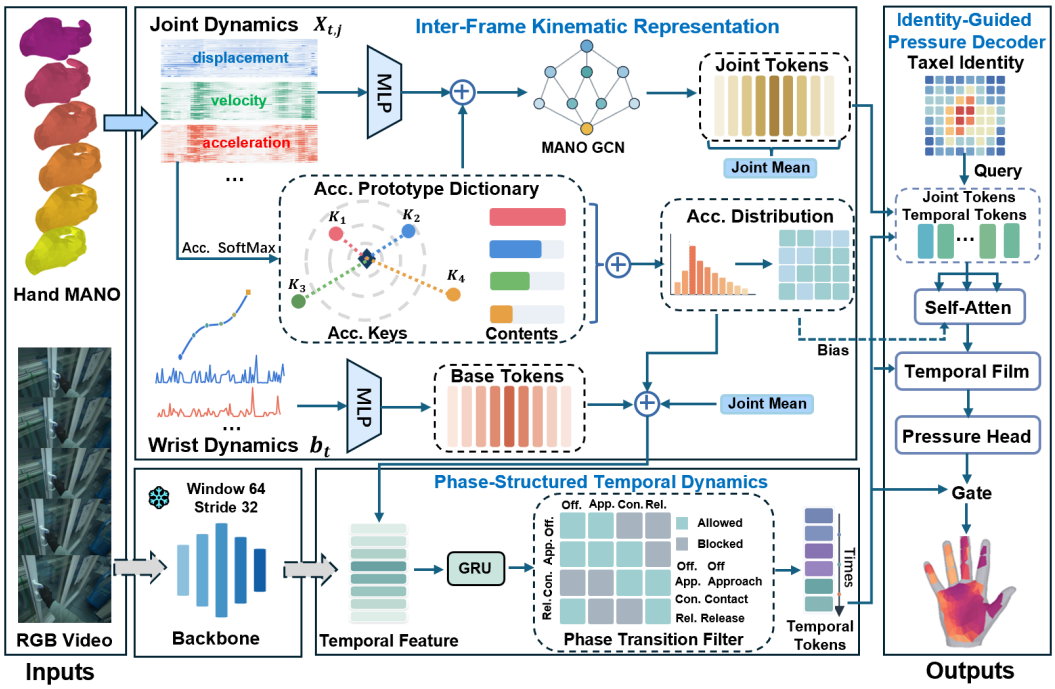


Figure 1: Overview of KIPS. IKR encodes wrist and joint kinematics into structured motion tokens. PSTD models contact-phase dynamics with optional RGB context. IGPD uses location-specific queries to decode dense pressure.

Acceleration prototype dictionary. To capture recurring joint acceleration patterns, we learn a key-content dictionary with $K = 16$ entries.

$$\mathcal{P} = \{(\mathbf{c}_k, \mathbf{e}_k)\}_{k=1}^{K}, \qquad \mathbf{c}_k \in \mathbb{R}^3, \quad \mathbf{e}_k \in \mathbb{R}^d, \tag{5}$$

where $\mathbf{c}_k$ is a key in acceleration space and $\mathbf{e}_k$ is its associated content, both are jointly learned and shared across samples. The key encodes a recurring acceleration pattern, while the content stores a reusable latent feature. Each joint is softly assigned to the dictionary according to

$$w_{t,j,k} = \frac{\exp\left(-\|\mathbf{a}_{t,j} - \mathbf{c}_k\|_2^2/\tau\right)}{\sum_{k'=1}^{K} \exp(-\|\mathbf{a}_{t,j} - \mathbf{c}_{k'}\|_2^2/\tau)}, \tag{6}$$

where $\tau = 0.3$ controls the temperature and $\mathbf{w}_{t,j}$ defines the distribution over prototypes. The projected joint token fuses the learned joint identity with the retrieved prototype content

$$\mathbf{J}_{t,j}^{(0)} = \phi_{\mathrm{x}}(\mathbf{x}_{t,j}) + \mathbf{e}_j^{\mathrm{id}} + \sum_{k=1}^{K} w_{t,j,k}\mathbf{e}_k, \tag{7}$$

where ϕ_{x} is a shared MLP and $\mathbf{e}_j^{\mathrm{id}} \in \mathbb{R}^d$ is a learnable identity for joint j. These fused tokens are then propagated over the MANO graph to exchange information among connected joints. For relation types r including self, parent, and child connections, each layer updates joint token as

$$\mathbf{J}_{t,u}^{(\ell+1)} = \mathbf{J}_{t,u}^{(\ell)} + \rho\left(\mathrm{LN}\left(\sum_{r\in\mathcal{R}} \frac{1}{|\mathcal{N}_r(u)|} \sum_{v\in\mathcal{N}_r(u)} W_r^{(\ell)}\mathbf{J}_{t,v}^{(\ell)}\right)\right), \tag{8}$$

where $\mathcal{N}_r(u)$ is the neighbor set of joint u under relation r, $W_r^{(\ell)}$ is a relation-specific linear map, and ρ is a GELU activation. We stack $L = 2$ layers in our implementation.

Averaging the assignments over joints gives a frame-level prototype distribution:

$$\bar{\mathbf{w}}_t = \frac{1}{J}\sum_{j=1}^{J} \mathbf{w}_{t,j} \in \mathbb{R}^K. \tag{9}$$

The IKR output consists of graph-encoded joint tokens $\{\mathbf{J}_{t,j}\}$, their mean-pooled summary $\bar{\mathbf{J}}_t$ a base token $\mathbf{B}_t$ from the wrist observation, and a spatial prior $\mathbf{s}_t$ obtained by projecting the frame-level prototype distribution onto the output layout via a low-rank prototype-to-location basis.

3.2 Phase-Structured Temporal Dynamics

Grasp pressure evolves with contact progression, emerging at contact onset, persisting during sustained contact, and decaying through release. PSTD models this phase structure while optionally incorporating RGB appearance as complementary context.

Temporal state encoding. We construct the frame-level representation by combining joint averaged token $\bar{\mathbf{J}}_t$, global wrist motion $\mathbf{B}_t$, and an encoding of time $\mathrm{PE}(t)$ as

$$\mathbf{f}_t = \bar{\mathbf{J}}_t + \mathbf{B}_t + \mathbb{I}_{\mathrm{rgb}}\phi_{\mathrm{rgb}}(\mathbf{r}_t) + \mathrm{PE}(t), \tag{10}$$

where $\mathbf{r}_t$ is an ℓ_2 normalized feature extracted by a frozen ConvNeXt (Liu et al., 2022; Woo et al., 2023) encoder. The indicator $\mathbb{I}_{\mathrm{rgb}}$ is zero for the kinematics only model. A bidirectional GRU (Cho et al., 2014) then aggregates the sequence as

$$\mathbf{h}_t = W_h \left[\overrightarrow{\mathrm{GRU}}(\mathbf{f}_{1:t}), \overleftarrow{\mathrm{GRU}}(\mathbf{f}_{t:T})\right]. \tag{11}$$

Structured contact filtering. We represent the interaction using four phases from the state set $\mathcal{S} = \{\mathrm{Off}, \mathrm{Approach}, \mathrm{Contact}, \mathrm{Release}\}$. An emission distribution is predicted from the temporal state and acceleration distribution as

$$\mathbf{e}_t = \mathrm{softmax}\left(W_e[\mathbf{h}_t, \bar{\mathbf{w}}_t]\right). \tag{12}$$

We parameterize phase transitions with learnable logits $A \in \mathbb{R}^{4\times4}$ and mask invalid edges using a valid transition set $\mathcal{E}$. Self-transitions are permitted for every phase. The allowed cross-phase transitions are (Off $\leftrightarrow$ Approach, Approach $\rightarrow$ Contact, Contact $\rightarrow$ Release, Release $\rightarrow$ Off). Direct jump from Off to Contact is forbidden. The normalized transition probability is

$$A_{ss'} = \frac{\mathbb{I}[(s,s') \in \mathcal{E}] \exp(\widetilde{A}_{ss'})}{\sum_{u\in\mathcal{S}} \mathbb{I}[(s,u) \in \mathcal{E}] \exp(\widetilde{A}_{su})}. \tag{13}$$

A differentiable forward filter combines the current emission with the preceding state as

$$\log \widetilde{\alpha}_t(s) = \log e_t(s) + \mathrm{LSE}_{s'}\big(\log \alpha_{t-1}(s') + \log A_{s's}\big), \qquad \boldsymbol{\pi}_t = \boldsymbol{\alpha}_t = \mathrm{softmax}\big(\log \widetilde{\boldsymbol{\alpha}}_t\big), \tag{14}$$

where $\boldsymbol{\pi}_t$ is the filtered phase posterior and LSE denotes log-sum exponential aggregation. This recursion combines the current emission with the transition-consistent posterior from the previous frame, suppressing temporally invalid phase changes and producing a frame-level pressure gate:

$$g_t = \sum_{s\in\mathcal{S}} \pi_t(s)\, \sigma(\omega_s)\, \bar{g}_s, \qquad \bar{\mathbf{g}} = (0.01, 0.05, 1.0, 0.05), \tag{15}$$

where ω_s is a learnable phase-specific strength. The fixed coefficients set **Off** to near zero, attenuate the responses during **Approach** and **Release**, and retain the response during **Contact**.

3.3 Identity-Guided Pressure Decoder

IGPD maps kinematic evidence to location-specific pressure estimates. Each location i, corresponding to a taxel or MANO vertex, is represented by a learnable identity query $\mathbf{q}_i$. At frame t, each query attends to the 21 graph-encoded joint tokens and a temporal token derived from $\mathbf{h}_t$.

Anatomically constrained decoding. Each query attends to an anatomically defined token subset $\mathcal{A}_i$. A finger query accesses its kinematic chain together with the wrist and temporal tokens, whereas a palm query accesses all joint and temporal tokens. For head m, the masked attention weights are

$$\alpha^{(m)}_{t,i,u} = \frac{\mathbb{I}[u \in \mathcal{A}_i] \exp\left(\langle W_Q^{(m)}\mathbf{q}_i, W_K^{(m)}\mathbf{k}_{t,u}\rangle / \sqrt{d_m}\right)}{\sum_{v\in\mathcal{A}_i} \exp\left(\langle W_Q^{(m)}\mathbf{q}_i, W_K^{(m)}\mathbf{k}_{t,v}\rangle / \sqrt{d_m}\right)}. \tag{16}$$

The outputs of all heads are concatenated to form

$$\mathbf{z}_{t,i} = W_{out}\,\mathrm{Concat}_m\left(\sum_{u\in\mathcal{A}_i}\alpha_{t,i,u}^{(m)}W_V^{(m)}\mathbf{k}_{t,u}\right). \tag{17}$$

The temporal state modulates each pressure representation through feature-wise affine conditioning:

$$\widetilde{\mathbf{z}}_{t,i} = \big(1+\gamma_i(\mathbf{h}_t)\big)\mathbf{z}_{t,i} + \beta_i(\mathbf{h}_t). \tag{18}$$

The ungated pressure magnitude and contact probability are predicted as

$$m_{t,i} = \mathrm{softplus}\left(\phi_p(\widetilde{\mathbf{z}}_{t,i}) + \lambda_s s_{t,i} + r_i(\mathbf{h}_t)\right), \qquad c_{t,i} = \sigma\left(\phi_c(\widetilde{\mathbf{z}}_{t,i}) + \boldsymbol{\pi}_t^\top \mathbf{b}_c\right). \tag{19}$$

The residual $r_i(\mathbf{h}_t)$ allows the shared temporal state to refine the candidate pressure magnitude. The final prediction factorizes into the magnitude $m_{t,i}$, frame-level phase gate g_t, and local contact probability $c_{t,i}$, corresponding respectively to how much pressure is present, when it is active, and where contact occurs. The final pressure response is

$$\widehat{p}_{t,i} = m_{t,i}\, g_t\, c_{t,i}. \tag{20}$$

3.4 Training Objective

We define the binary contact label as $y_{t,i}^c = \mathbb{I}[p_{t,i} > \eta]$, where η is the contact threshold.

Pressure regression. We supervise the pressure magnitude over the complete output using

$$\mathcal{L}_{\text{full}} = \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\left|\widehat{p}_{t,i} - p_{t,i}\right|. \tag{21}$$

Since inactive regions dominate the output field, we additionally compute the metric within measured contact regions to directly assess pressure estimation on active contacts.

$$\mathcal{L}_{\text{active}} = \frac{\sum_{t,i} y_{t,i}^c\,|\widehat{p}_{t,i} - p_{t,i}|}{\sum_{t,i} y_{t,i}^c + \epsilon}. \tag{22}$$

Contact supervision. The model predicts an explicit contact probability $c_{t,i}$ for each taxel or vertex. We supervise this output using focal loss (Lin et al., 2017):

$$\mathcal{L}_{\text{taxel}} = -\frac{1}{TN}\sum_{t,i}\Big[\alpha y_{t,i}^c(1-c_{t,i})^\gamma \log c_{t,i} + (1-\alpha)(1-y_{t,i}^c)c_{t,i}^\gamma \log(1-c_{t,i})\Big]. \tag{23}$$

This term reduces the effect of the imbalance between contact and non-contact taxels. To align regressed pressure with the contact threshold, we convert the signed pressure margin into a contact probability and supervise it with binary cross-entropy:

$$\widetilde{c}_{t,i} = \sigma(\widehat{p}_{t,i} - \eta), \quad \mathcal{L}_{\text{margin}} = -\frac{1}{TN}\sum_{t,i}\big[y_{t,i}^c \log \widetilde{c}_{t,i} + (1-y_{t,i}^c)\log(1-\widetilde{c}_{t,i})\big]. \tag{24}$$

Phase supervision. We supervise the filtered posterior $\boldsymbol{\pi}_t$ with cross-entropy against phase labels y_t^ϕ derived from the measured contact sequence, yielding

$$\mathcal{L}_{\text{phase}} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}}\mathbb{I}\Big[y_t^\phi = s\Big]\log \pi_t(s). \tag{25}$$

This supervision anchors the structured temporal posterior to the interaction phases used by the phase-dependent pressure decoder.

Complete objective. The overall training objective is formulated as follows:

$$\mathcal{L} = \lambda_f\mathcal{L}_{\text{full}} + \lambda_a\mathcal{L}_{\text{active}} + \lambda_t\mathcal{L}_{\text{taxel}} + \lambda_m\mathcal{L}_{\text{margin}} + \lambda_p\mathcal{L}_{\text{phase}}, \tag{26}$$

which we empirically set to $\lambda_f = 1$, $\lambda_a = 0.25$, $\lambda_t = 0.1$, $\lambda_m = 0.2$, and $\lambda_p = 0.1$.

Table 1: Results on the EgoTactile gloved splits. **Best** and second best results are highlighted.

			Object held-out				Subject held-out			
Input	Method	Venue	MAE↓	CoP↓	V-IoU↑	C-IoU↑	MAE↓	CoP↓	V-IoU↑	C-IoU↑
RGB	PressureVision	ECCV 2022	9.200	12.500	0.168	0.245	10.500	14.200	0.143	0.212
RGB	EgoPressureFormer	ICML 2026	6.200	7.500	0.265	0.368	7.500	8.800	0.228	0.325
RGB	EgoPressureDiff	ICML 2026	3.400	3.100	0.389	0.563	5.800	4.700	0.340	0.512
Kinematics	KIPS-K	Ours	2.639	2.946	0.460	0.546	3.021	**2.516**	0.413	0.453
Kinematics + RGB	KIPS-K+RGB	Ours	**2.302**	**2.491**	**0.524**	**0.604**	**2.569**	2.521	**0.461**	**0.517**

4 EXPERIMENTS

We investigate how well hand kinematics can predict dense grasp pressure and under what conditions it is most informative. KIPS-K uses hand kinematics as the sole input to the pressure estimator, whereas KIPS-K+RGB additionally conditions the temporal state on frozen RGB features. Both variants use identical MANO trajectories recovered by the same upstream pose estimator.

4.1 EXPERIMENTAL SETUP

Datasets and metrics. We evaluate on EgoTactile (Zeng et al., 2026) and OpenTouch (Song et al., 2025). EgoTactile contains 319k egocentric video frames aligned with a 256-taxel glove across 768 clips. We follow its official object-held-out, subject-held-out, and bare-hand transfer protocols. OpenTouch comprises 2,900 in-the-wild clips (5.1 hours) with synchronized video, pose, and full-hand pressure. We use its official split of 592 test clips. We report clip-level MAE, CoP, volumetric V-IoU, and contact C-IoU on EgoTactile, and vertex MAE, RMSE, NRMSE, and CoP on OpenTouch. Metric computation details are provided in Appendix A.3.

Baselines. We compare against PressureVision (Grady et al., 2022), EgoPressureFormer, and EgoPressureDiff under their published protocols (Zeng et al., 2026). We further initialize HACO (Jung & Lee, 2025) from its released checkpoint, attach a pressure estimation head, and fine-tune it on each benchmark. Under the EgoTactile protocol, a taxel is in contact if its pressure exceeds $5\,\mathrm{N}$. Temporally, a finger is active if at least three of its taxels are active, and a frame is active if any finger is active or at least five taxels are active overall.

Implementation details. KIPS uses 64-frame windows with stride 32 and a hidden dimension of 128. The architecture comprises two MANO graph layers, four attention heads, 16 acceleration prototypes, a rank-16 taxel basis, and a one-layer temporal GRU. We apply dropout of 0.1 and normalize motion channels using training-set statistics. Training uses AdamW (Loshchilov & Hutter, 2017) for 30 epochs with a learning rate of 3×10^{-4}, weight decay 10^{-4}, batch size 64, and gradient norm clipping at 1.0. RGB features are extracted by a frozen ConvNeXt-Base encoder. All experiments use Python 3.11 and PyTorch 2.6 with CUDA 12.4. More details can be found in Appendix.

4.2 MAIN RESULTS

EgoTactile gloved evaluation. Table 1 shows that KIPS-K, using hand kinematics alone in the pressure estimator, outperforms every RGB-based baseline on six of eight metrics. Relative to EgoPressureDiff, it reduces MAE by 22.4% and 47.9%, lowers CoP error by 5.0% and 46.5%, and improves V-IoU by 18.3% and 21.5% on the object- and subject-held-out splits, respectively. These gains establish hand kinematics as a strong primary signal for dense pressure estimation. Adding RGB yields the best performance on seven of eight metrics, including all four object-held-out measures. Compared with KIPS-K, KIPS-K+RGB further reduces MAE by 12.8% and 15.0%, improves V-IoU by 13.9% and 11.6%, and raises C-IoU by 10.6% and 14.1% across the two splits. The consistent overlap gains indicate that appearance is especially useful for refining spatial pressure support left ambiguous by motion. KIPS establishes a new state of the art on EgoTactile.

Bare hand transfer. Table 2 evaluates gloved-to-bare-hand transfer. Without adaptation, KIPS-K and KIPS-K+RGB outperform appearance-based methods in pressure magnitude and load localization, showing that kinematics remains informative under appearance and sensing shifts. Appearance-based methods still lead in spatial overlap, indicating that zero-shot contact support is more domain-

Table 2: Results on the EgoTactile bare splits. **Best** and second best results are highlighted.

			zero-shot				fine tune			
Input	Method	Venue	MAE↓	CoP↓	V-IoU↑	C-IoU↑	MAE↓	CoP↓	V-IoU↑	C-IoU↑
RGB	PressureVision	ECCV 2022	11.500	16.000	0.102	0.155	9.500	12.800	0.150	0.228
RGB	EgoPressureFormer	ICML 2026	8.800	10.500	0.185	0.248	6.400	7.800	0.258	0.362
RGB	EgoPressureDiff	ICML 2026	4.500	4.200	**0.325**	**0.472**	3.200	2.900	0.402	0.578
Kinematics	KIPS-K	Ours	3.502	3.234	0.310	0.423	2.330	2.574	0.529	0.598
Kinematics + RGB	KIPS-K+RGB	Ours	**3.108**	**2.482**	0.312	0.409	**2.153**	**2.572**	**0.563**	**0.620**

Table 3: Results on the OpenTouch. **Best** and second best results are highlighted.

Input	Method	Venue	MAE↓ (kPa)	RMSE↓ (kPa)	NRMSE↓ (%)	CoP↓ (mm)
RGB	HACO	NeurIPS 2025	1.659	4.387	9.16	1.390
Kinematics	KIPS-K	Ours	1.420	3.760	7.85	1.193
Kinematics + RGB	KIPS-K+RGB	Ours	**1.369**	**3.668**	**7.66**	**1.162**

sensitive. After fine-tuning, both KIPS-K variants surpass all baselines. These results establish kinematics as a transferable structural signal that appearance and adaptation further refine.

OpenTouch. Table 3 further confirms the advantage of kinematics on OpenTouch. KIPS-K reduces MAE and RMSE relative to HACO by approximately 14%, while RGB conditioning provides a further gain of approximately 3.6%. This consistent improvement demonstrates that the benefits of kinematic evidence generalize across datasets.

Figure 2 compares regional localization across six hand regions. Scores are normalized within each region and log-compressed for visualization. KIPS-K consistently outperforms KIPS-RGB and closely tracks KIPS-K+RGB, showing that kinematics captures most of the localization signal, with RGB providing complementary refinements. The remaining gap is concentrated around the little finger, which is the most difficult region for all methods.

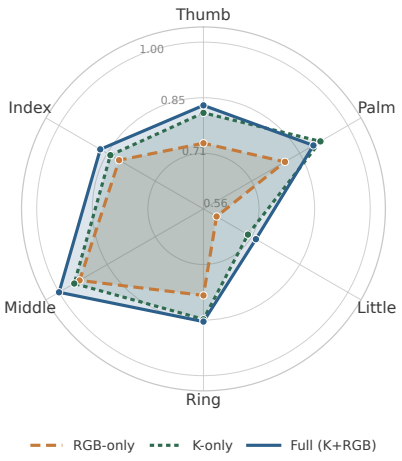


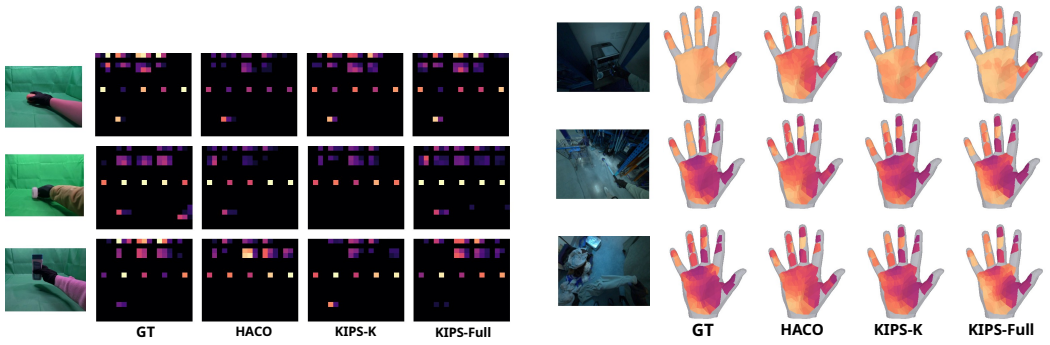
Figure 2: Region-wise CoP on OpenTouch. Larger is better.

4.3 Qualitative Results

Figure 3 compares KIPS with HACO and the ground truth on both datasets. EgoTactile pressure is displayed on the glove sensor grid, while OpenTouch predictions are projected onto the MANO surface. KIPS-K recovers the dominant contact regions and overall pressure structure from kinematics alone. RGB conditioning further refines local pressure magnitudes and spatial support, bringing the predictions closer to the ground truth for both concentrated fingertip contacts and distributed pressure patterns. The resulting predictions more closely follow both concentrated fingertip responses and distributed pressure patterns than HACO. Some discrepancies remain for weak precise pressure magnitudes.

4.4 When does appearance help?

We stratify contact frames into eight equal-frequency bins by wrist acceleration magnitude. Figure 4 reveals two consistent trends. First, the MAE of KIPS-K decreases toward higher-motion bins on both datasets, indicating more accurate pressure estimation under stronger motion. Second, the relative MAE reduction provided by RGB narrows from 12.2% to 5.7% on EgoTactile and from 4.6% to 1.3% on OpenTouch. The same pattern holds for CoP, with the RGB gain approaching zero in the highest-motion OpenTouch bin. These results suggest that appearance contributes most in low-motion regimes, while its marginal benefit diminishes as motion strengthens. Appendix B.1



(a) EgoTactile **(b) OpenTouch**

Figure 3: Qualitative results on EgoTactile and OpenTouch.

supports a negative association between acceleration and prediction error after accounting for pressure magnitude and clip-level differences. These trend are consistent with the intuition that stronger grasp motion provides more informative kinematic cues.

4.5 Ablation Studies

Figure 5 evaluates the principal components of KIPS on OpenTouch. Removing joint-resolved dynamics produces the largest degradation, increasing MAE by 18.1% and RMSE by 17.2%. Replacing temporal aggregation with frame-wise prediction yields the second-largest degradation, with MAE and RMSE increasing by 14.0% and 10.0%, respectively. This indicates that isolated poses capture contact evolution less effectively than temporal representations. Using joint dynamics alone also under-performs the full model, which indicates that wrist transport provides complementary information about global hand and object motion. Phase dynamics, acceleration prototypes, transition filtering, and explicit acceleration produce smaller but consistent improvements. Together, these results validate the central design choices of KIPS.

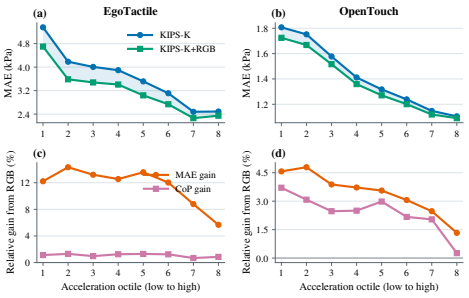


Figure 4: Results across equal frequency acceleration bins. **(a,b)**: MAE. **(c,d)**: Relative gains from RGB appearance.

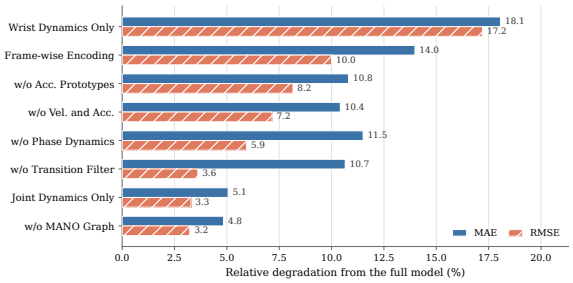


Figure 5: Relative degradation on OpenTouch after removing or replacing individual components.

5 Conclusion

We introduced KIPS, a kinematics-led framework that treats hand motion as an implicit physical sensor for dense egocentric grasp pressure estimation. KIPS combines multi-granularity kinematic encoding, phase-structured temporal modeling, and identity-guided pressure decoding. Across EgoTactile and OpenTouch, the kinematics-only model outperforms RGB-based methods on most metrics, while appearance conditioning provides complementary gains. Motion-stratified analysis further shows that the marginal benefit of RGB diminishes as motion strengthens. Together, these results establish hand kinematics as a strong primary modality for dense pressure estimation in everyday grasping across diverse interaction and sensing conditions.

AI USE STATEMENT

AI was used to assist with language polishing and improve the clarity of the manuscript. The tool was not used to generate experimental results or scientific claims. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work does not involve the collection of new data or the release of new datasets. All experiments are conducted on existing hand-object grasp benchmarks and focus on model training, analysis, and evaluation. We do not identify ethical risks beyond those associated with the use of the underlying benchmark data and models. The authors have reviewed the experimental design and reporting to ensure that the results and conclusions are presented accurately and responsibly.

REPRODUCIBILITY STATEMENT

We provide the main implementation and experimental details required to reproduce KIPS in the paper and Appendix. Datasets split, contact thresholds, and valid evaluation locations follow the corresponding benchmark protocols. The complete project code will be released upon acceptance.

REFERENCES

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7061–7071, 2025.

Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8701–8711, 2019.

Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *European Conference on Computer Vision (ECCV)*, pp. 361–378. Springer, 2020.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. Dexycb: A benchmark for capturing hand grasping of objects. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9040–9049, 2021.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734. Association for Computational Linguistics, 2014.

Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20545–20554, 2022.

Jeremy A. Collins, Cody Houff, Patrick Grady, and Charles C. Kemp. Visual contact pressure estimation for grippers in the wild. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10947–10954, 2023.

Enric Corona, Albert Pumarola, Guillem Alenyà, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5030–5040, 2020.

Kiana Ehsani, Shubham Tulsiani, Saurabh Gupta, Ali Farhadi, and Abhinav Gupta. Use the force, luke! learning to predict physical forces by simulating effects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 221–230, 2020.

J Randall Flanagan and Alan M Wing. Modulation of grip force with load force during point-to-point arm movements. *Experimental brain research*, 95(1):131–143, 1993.

Fan Gao, Mark L Latash, and Vladimir M Zatsiorsky. Internal forces during object manipulation. *Experimental brain research*, 165(1):69–83, 2005.

Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. Contactopt: Optimizing contact to improve grasps. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1471–1481, 2021.

Patrick Grady, Chengcheng Tang, Samarth Brahmbhatt, Christopher D Twigg, Chengde Wan, James Hays, and Charles C Kemp. Pressurevision: estimating hand pressure from a single rgb image. In *European Conference on Computer Vision (ECCV)*, pp. 328–345. Springer, 2022.

Patrick Grady, Jeremy A. Collins, Chengcheng Tang, Christopher D. Twigg, Kunal Aneja, James Hays, and Charles C. Kemp. Pressurevision++: Estimating fingertip pressure from diverse rgb images. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8683–8693, 2024.

Steven Henderson and Steven Feiner. Opportunistic tangible user interfaces for augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 16(1):4–16, 2010.

Subin Jeon, Byungjun Kim, and Hanbyul Joo. Hope: Hand-object pressure estimation from monocular videos. *arXiv preprint arXiv:2608.06192*, 2026.

Daniel Sungho Jung and Kyoung Mu Lee. Learning dense hand contact estimation from imbalanced data. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Zongmian Li, Jiri Sedlar, Justin Carpentier, Ivan Laptev, Nicolas Mansard, and Josef Sivic. Estimating 3d motion and forces of person-object interactions from monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8632–8641, 2019.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection . In *IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11966–11976. IEEE, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Priyanka Mandikal and Kristen Grauman. Learning dexterous grasping with object-centric visual affordances. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6169–6176, 2021.

Manuel Meier, Paul Streli, Andreas Fender, and Christian Holz. Tapid: Rapid touch interaction in virtual reality using wearable sensing. In *IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 519–528, 2021.

Huiming Pan, Kezhe Zhu, Dongxuan Li, Yueyuan Chen, Shuo Jiang, Bin He, and Peter B. Shull. Decoding human touch noninvasively: Tactile inference from emg and kinematics using aettacnet. *IEEE Transactions on Industrial Informatics*, 22(6):5117–5128, 2026.

Tu-Hoa Pham, Nikolaos Kyriazis, Antonis A. Argyros, and Abderrahmane Kheddar. Hand-object contact force estimation from markerless visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2883–2896, 2018.

Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *European conference on computer vision (ECCV)*, pp. 71–87. Springer, 2020.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 36(6), 2017.

Kyungjin Seo, Junghoon Seo, Hanseok Jeong, Sangpil Kim, and Sang Ho Yoon. Posture-informed muscular force learning for robust hand pressure estimation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Yuxin Ray Song, Jinzhou Li, Rao Fu, Devin Murphy, Kaichen Zhou, Rishi Shiv, Yaqi Li, Haoyu Xiong, Crystal Elaine Owens, Yilun Du, et al. Opentouch: Bringing full-hand touch to real-world interaction. *arXiv preprint arXiv:2512.16842*, 2025.

Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, 569 (7758):698–702, 2019.

Yinqiao Wang, Hao Xu, Pheng-Ann Heng, and Chi-Wing Fu. Unihope: A unified approach for hand-only and hand-object pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12231–12241, 2025.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16133–16142, 2023.

Yuan Zeng, Yujia Shi, Tiao Tan, Xingting Li, Yaqi Qin, Zongqing Lu, Wenming Yang, Jing-Hao Xue, and Qingmin Liao. Egotactile: Learning grasp pressure for everyday objects from egocentric video. *arXiv preprint arXiv:2606.09243*, 2026.

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1805–1815, 2025.

Yufei Zhang, Jeffrey O. Kephart, Zijun Cui, and Qiang Ji. Physpt: Physics-aware pretrained transformer for estimating human dynamics from monocular videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2305–2317, 2024.

Yiming Zhao, Taein Kwon, Paul Streli, Marc Pollefeys, and Christian Holz. Egopressure: A dataset for hand pressure and pose estimation in egocentric vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27727–27738, June 2025.

APPENDIX-HAND KINEMATICS AS AN IMPLICIT PHYSICAL SENSOR FOR EGOCENTRIC GRASP PRESSURE ESTIMATION

A ADDITIONAL IMPLEMENTATION AND EVALUATION DETAILS

This section clarifies several implementation details that are not expanded in the main paper.

A.1 DESIGN CLARIFICATIONS

Soft prototype composition. The acceleration prototype dictionary provides a continuous representation of joint dynamics rather than a discrete assignment of motion states. An observed joint acceleration may be comparably close to multiple learned keys, particularly under gradual motion transitions or noise in the recovered hand trajectory. Hard assignment to a single nearest prototype would discard this ambiguity and introduce discontinuous feature changes near prototype boundaries. KIPS instead retrieves a similarity-weighted combination of prototype contents, allowing nearby acceleration patterns to induce smoothly varying motion representations.

The retrieved content is incorporated into each joint representation before propagation over the MANO graph. The resulting prototype-conditioned features therefore reach the pressure decoder through the graph-encoded joint tokens attended to by each output query. In parallel, the prototype assignments are averaged across joints to obtain a frame-level distribution, which conditions both the temporal phase emission and the prototype-conditioned spatial prior used by the pressure decoder. The dictionary should therefore be interpreted as a set of softly composable acceleration-conditioned motion components, rather than as discrete interaction phases or fixed pressure templates.

Prototype-conditioned spatial prior. In addition to the prototype content retrieved by each joint token, KIPS converts the frame-level prototype distribution into a lightweight bias over the output locations. Let $\bar{\mathbf{w}}_t \in \mathbb{R}^K$ denote the prototype assignment averaged over the J joints. We construct a non-negative low-rank basis as:

$$\widetilde{B} = \mathrm{softplus}(U)\,\mathrm{softplus}(V), \qquad U \in \mathbb{R}^{K\times r}, \qquad V \in \mathbb{R}^{r\times N}, \tag{27}$$

where $K = 16$ is the number of acceleration prototypes, $r = 16$ is the basis rank, and N is the number of output locations. The spatial prior provides an acceleration-conditioned additive bias to the decoder. We normalize each row to obtain:

$$B_{k,:} = \frac{\widetilde{B}_{k,:}}{\|\widetilde{B}_{k,:}\|_1 + \epsilon}. \tag{28}$$

The prototype-conditioned spatial prior is then computed as:

$$\mathbf{s}_t = \bar{\mathbf{w}}_t B \in \mathbb{R}^N. \tag{29}$$

The resulting value $s_{t,i}$ is added to the pressure-magnitude logit at output location i:

$$m_{t,i} = \mathrm{softplus}\left(\phi_p(\widetilde{\mathbf{z}}_{t,i}) + \lambda_s s_{t,i} + r_i(\mathbf{h}_t)\right). \tag{30}$$

A.2 TARGETS AND SEQUENCE RECONSTRUCTION

Phase-target construction. The phase targets are deterministically derived from the measured hand-contact sequence. For EgoTactile, taxel i is active at frame t when its measured pressure exceeds $5\,\mathrm{N}$. A finger is active when at least three of its taxels are active. The hand-contact flag y_t^h is positive when any finger is active or when at least five taxels are active across the complete hand.

Frames with $y_t^h = 1$ are assigned to CONTACT. The eight frames immediately preceding each contact onset are assigned to APPROACH. The eight frames immediately following each contact

offset are assigned to RELEASE. All remaining frames are assigned to OFF. At 30 fps, the approach and release intervals each cover approximately 267 ms. Intervals are truncated at sequence boundaries when fewer than eight frames are available. These targets supervise the phase posterior during training. At inference time, the model operates only on its predicted emission probabilities and the recurrent transition filter.

Full-sequence reconstruction. KIPS processes overlapping temporal windows. Let Ω_t denote the set of windows containing frame t. The final full-sequence prediction is

$$\widehat{\mathbf{p}}_t = \frac{1}{|\Omega_t|} \sum_{w \in \Omega_t} \widehat{\mathbf{p}}_t^{(w)}. \tag{31}$$

Overlapping estimates are averaged before evaluation. This produces one pressure field for every original video frame and avoids discontinuities at window boundaries.

A.3 EVALUATION MEASURES

Let $\mathcal{I}$ denote the set of valid taxels or vertices. Let $p_{t,i}$ and $\widehat{p}_{t,i}$ denote the measured and predicted pressure at frame t and location i.

Mean absolute error. MAE measures the average pressure-magnitude error:

$$\mathrm{MAE} = \frac{1}{T|\mathcal{I}|} \sum_{t=1}^{T} \sum_{i \in \mathcal{I}} \left| \widehat{p}_{t,i} - p_{t,i} \right|. \tag{32}$$

Center-of-pressure error. Let $\{\mathcal{R}_r\}_{r=1}^{6}$ denote the sets of valid taxels associated with the five fingertips and the palm, and let $\mathbf{c}_i$ denote the fitted-grid coordinate of taxel i. For anatomical region r at frame t, the measured and predicted pressure centroids are defined as

$$\boldsymbol{\mu}^{(r)}(\mathbf{p}_t) = \frac{\sum_{i \in \mathcal{R}_r} p_{t,i} \mathbf{c}_i}{\sum_{i \in \mathcal{R}_r} p_{t,i} + \epsilon}, \qquad \boldsymbol{\mu}^{(r)}(\widehat{\mathbf{p}}_t) = \frac{\sum_{i \in \mathcal{R}_r} \widehat{p}_{t,i} \mathbf{c}_i}{\sum_{i \in \mathcal{R}_r} \widehat{p}_{t,i} + \epsilon}, \tag{33}$$

The CoP error for region r is the average Euclidean distance between the predicted and measured pressure over its valid frames:

$$\mathrm{CoP}^{(r)} = \frac{1}{|\mathcal{T}_r|} \sum_{t \in \mathcal{T}_r} \left\| \boldsymbol{\mu}^{(r)}(\widehat{\mathbf{p}}_t) - \boldsymbol{\mu}^{(r)}(\mathbf{p}_t) \right\|_2. \tag{34}$$

The final CoP error is the unweighted mean over the six anatomical regions:

$$\mathrm{CoP} = \frac{1}{6} \sum_{r=1}^{6} \mathrm{CoP}^{(r)}. \tag{35}$$

The metric is reported in sensor-layout grid units and measures whether the predicted pressure is localized correctly within each anatomical region. Including frames activated by either prediction or ground truth ensures that spatially misplaced false-positive contacts are also penalized.

Volumetric pressure overlap. V-IoU measures the overlap between the continuous predicted and measured pressure fields:

$$\text{V-IoU}_t = \frac{\sum_{i \in \mathcal{I}} \min(\widehat{p}_{t,i}, p_{t,i})}{\sum_{i \in \mathcal{I}} \max(\widehat{p}_{t,i}, p_{t,i}) + \epsilon}. \tag{36}$$

It jointly evaluates pressure magnitude and spatial support. A high V-IoU requires the two pressure fields to assign similar pressure values to similar hand locations.

Contact-region overlap. C-IoU evaluates binary contact geometry. Given the benchmark contact threshold η, we define

$$\mathcal{C}_t = \{i \in \mathcal{I} \mid p_{t,i} > \eta\}, \qquad \widehat{\mathcal{C}}_t = \{i \in \mathcal{I} \mid \widehat{p}_{t,i} > \eta\}. \tag{37}$$

The contact overlap C-IoU is

$$\text{C-IoU}_t = \frac{|\widehat{\mathcal{C}}_t \cap \mathcal{C}_t|}{|\widehat{\mathcal{C}}_t \cup \mathcal{C}_t| + \epsilon}. \tag{38}$$

Unlike V-IoU, C-IoU evaluates which regions are in contact without requiring their pressure magnitudes to match exactly.

Root mean squared error. RMSE measures the pressure-magnitude error while assigning greater weight to large deviations:

$$\text{RMSE} = \sqrt{\frac{1}{T|\mathcal{I}|} \sum_{t=1}^{T} \sum_{i \in \mathcal{I}} (\widehat{p}_{t,i} - p_{t,i})^2}. \tag{39}$$

RMSE is reported in the corresponding physical pressure unit and complements MAE by being more sensitive to large prediction errors.

B Additional Experimental Analysis

This section presents additional evaluations of KIPS from three perspectives: performance on an additional contact benchmark, analysis of its kinematic and temporal representations, and robustness and deployment characteristics. Unless otherwise stated, experiments on EgoTactile follow the official object-held-out gloved protocol.

B.1 Physical Evidence in Hand Kinematics

KIPS is motivated by the observation that hand motion contains information about the evolution of the contact load even when the hand-object interface is not directly visible. We further examine this relationship using the measured pressure and the recovered kinematic trajectory, independently of the pressure prediction model. For every frame, we summarize joint displacement, velocity, and acceleration by their mean magnitude over the 21 MANO joints. The pressure response is summarized over the glove taxels.

Temporal association. To examine whether the dynamic kinematic channels covary with the measured contact state, we reduce each kinematic order to its mean joint-wise magnitude and sum the ground-truth taxel forces to obtain the aggregate contact force F_t. For each clip and temporal offset Δ, we compute $r_c(\Delta) = \text{corr}(k_c(t), F_c(t + \Delta))$ and aggregate the clip-level coefficients using Fisher's z transform. As shown in Figure 6(a), velocity and acceleration exhibit substantially stronger association with pressure than displacement. Their maximum correlation magnitudes are 0.355 and 0.297, respectively, and both peak at a positive lag within two frames. At the frame rate of 30 fps, this places their strongest association within approximately 60 ms of the subsequent pressure response. The correlation remains near its maximum over a short neighboring interval, indicating that the dynamic evidence is temporally local rather than tied to a single isolated sample.

Complementary dynamic information. We next examine whether the three kinematic orders retain distinct pressure-related information after accounting for their shared linear components. For each order, we regress both the kinematic signal and the aggregate contact force on the other two orders, and compute the correlation between the resulting residuals. Figure 6(b) reports the partial correlations over all frames and over contact frames only. The three orders retain distinct residual associations with the aggregate force: displacement is positively associated with the loading state,

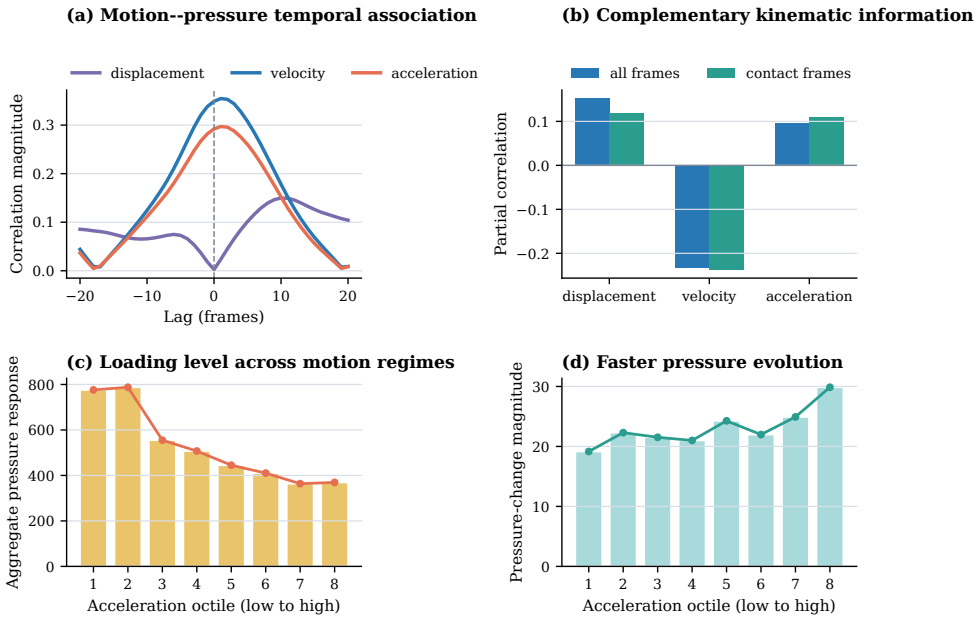


Figure 6: Physical evidence in hand kinematics on EgoTactile. **(a)** Magnitude of the lagged correlation between the three kinematic orders and the aggregate pressure response. **(b)** Partial correlations after controlling for the remaining kinematic orders. **(c)** Aggregate pressure response across acceleration octiles. **(d)** Pressure-change magnitude across the same octiles. Dynamic frames exhibit faster pressure evolution, complementing the sustained loading signal present in lower-motion frames.

Table 4: Pressure-controlled analysis across acceleration regimes.

	Pressure-controlled evidence			Scale-normalized diagnostics	
Dataset	**Consistent strata**	**Adjusted MAE change (95% CI)**	**Within-clip effect**	**NMAE: low → high (change; 95% CI)**	**Skill over mean low → high**
EgoTactile	$\mathbf{5/5}$	$\mathbf{-28.9\%}$ $[\mathbf{-37.8}, \mathbf{-19.2}]$	$\mathbf{-0.062}$	$70.4\% \rightarrow 66.5\%$ -5.5%; $[-14.8, +5.6]$	$13.4\% \rightarrow 54.2\%$
OpenTouch	$\mathbf{5/5}$	$\mathbf{-2.4\%}$ $[\mathbf{-4.5}, \mathbf{-0.8}]$	$\mathbf{-0.022}$	$\mathbf{3.82\%} \rightarrow \mathbf{2.28\%}$ $\mathbf{-40.2\%}$; $[\mathbf{-47.9}, \mathbf{-31.5}]$	$16.8\% \rightarrow 36.4\%$

while velocity shows the largest association in magnitude, consistent with lower motion during sustained loading. Acceleration retains a positive partial correlation of 0.096 over all frames and 0.110 within contact, indicating that its pressure-related information is not fully explained by displacement and velocity. The persistence of these associations within contact frames supports the complementary use of multiple kinematic orders beyond simply separating contact from non-contact.

Motion regimes. We partition frames into eight equal-frequency bins according to acceleration magnitude. Figures 6(c,d) reveal a consistent division between sustained loading and dynamic change. The aggregate pressure response decreases from 776.1 in the lowest-acceleration bin to 369.3 in the highest-acceleration bin, with a trend correlation of -0.936. In contrast, the average pressure-change magnitude increases from 19.2 to 29.9 per frame, with a trend correlation of 0.827. Thus, high-acceleration frames are associated with faster pressure evolution, whereas strong sustained loads occur more frequently in comparatively stable motion regimes. Together, these results support the use of multiple derivative orders: displacement captures configuration, while velocity and acceleration characterize how the contact state evolves over time.

Pressure-controlled prediction accuracy. Raw MAE can vary with target pressure magnitude and we address this confound through three complementary controls, as shown in Table 4. First, frames are jointly stratified by pressure quintile and acceleration tertile. Higher acceleration is asso-

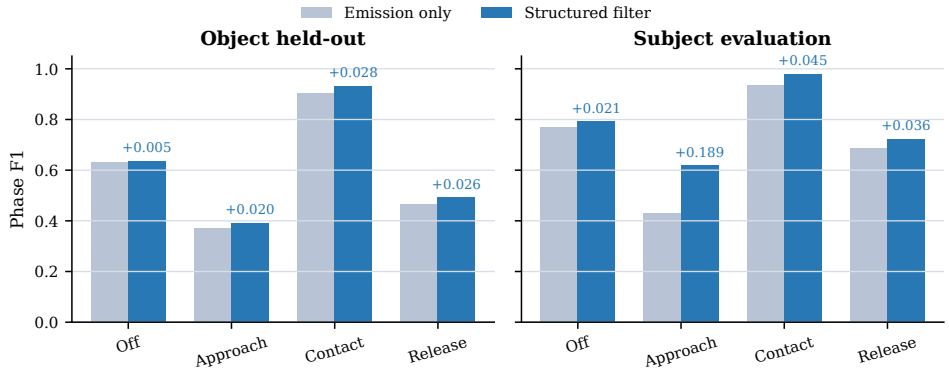


Figure 7: Phase prediction before and after structured temporal filtering. The filter improves all four phase F1 scores on the object-held-out evaluation and yields particularly strong gains for APPROACH and CONTACT on the subject evaluation.

ciated with lower KIPS-K error in all five pressure strata on both datasets. Second, we regress frame-level log MAE on normalized acceleration rank while controlling for pressure rank, its quadratic term, contact state, and clip fixed effects. The acceleration coefficient remains significantly negative on both EgoTactile and OpenTouch. Third, within-clip ranking produces the same negative association, reducing the influence of time-invariant object and subject differences. Confidence intervals are obtained from 2,000 clip-level bootstrap resamples. On OpenTouch, average pressure remains nearly unchanged from the lowest to the highest acceleration octile (47.34 versus 48.40 kPa), while NMAE decreases by 40.2%. On EgoTactile, the pressure-matched association is consistent across all five pressure strata and all five evaluated objects. Within-clip ranking yields the same negative association, with effects of -0.062 on EgoTactile and -0.022 on OpenTouch. These results indicate that the observed error reduction cannot be explained solely by differences in pressure magnitude and is consistent with stronger motion providing more informative kinematic evidence.

B.2 QUANTITATIVE EVALUATION OF PHASE-STRUCTURED DYNAMICS

The phase-structured temporal model represents an interaction using four states: OFF, APPROACH, CONTACT, and RELEASE. To quantify the contribution of structured filtering, we compare the final filtered posterior with the frame-wise emission distribution produced before recursive filtering. Phase targets are derived from the measured hand-contact sequence using the same deterministic phase construction as used during training.

Figure 7 shows that structured filtering improves the F1 score of every phase on the object-held-out evaluation. The improvement is particularly consistent for the three transition-related states: $+0.020$ for APPROACH, $+0.028$ for CONTACT, and $+0.026$ for RELEASE. On the subject evaluation, filtering increases APPROACH F1 from 0.428 to 0.617 and CONTACT F1 from 0.934 to 0.979. These gains show that the constrained temporal recurrence uses neighboring evidence to produce phase trajectories that are more coherent than independent per-frame emissions.

The benefit is also visible in event timing, as shown in Table 5. On the object-held-out split, the emission-only prediction activates contact 3.73 frames early on average, whereas the structured posterior yields an average onset error of 0.00 frames. On the subject evaluation, the onset error is reduced from -0.58 to -0.05 frames. This improved temporal localization directly supports the phase-dependent gate used by the pressure decoder.

B.3 MODALITY COMPARISON ON OPENTOUCH

Table 6 compares the RGB-only and kinematics-only variants under the same OpenTouch split and evaluation protocol. KIPS-K reduces MAE from 1.599 to 1.420 kPa relative to KIPS-RGB, with lower RMSE, NRMSE, and CoP error as well. This comparison shows that kinematic input is effective within KIPS and that RGB provides complementary information on OpenTouch.

Table 5: Contact-phase accuracy and onset localization. Negative onset values denote early activation relative to the pressure-derived contact sequence.

Evaluation	Inference	Contact F1	Onset (frames)
Object held-out	Emission only	0.905	-3.73
	Structured filter	**0.933**	0.00
Subject evaluation	Emission only	0.934	-0.58
	Structured filter	**0.979**	-0.05

Table 6: Controlled modality comparison on OpenTouch.

Input	**Variant**	**MAE↓** (kPa)	**RMSE↓** (kPa)	**NRMSE↓** (%)	**CoP↓** (mm)
RGB	KIPS-RGB	1.599	4.210	8.42	1.341
Kinematics	KIPS-K	1.420	3.760	7.85	1.193

B.4 Temporally Aligned Appearance Conditioning

KIPS uses appearance as complementary evidence while retaining a kinematics-led representation. We analyze whether this conditioning depends on the frame-wise alignment between visual and kinematic observations. Starting from the same trained KIPS-K+RGB model, we evaluate four visual streams: the original aligned features, a random temporal permutation within each window, repetition of the first-frame feature, and a zeroed visual stream.

Table 7 and Figure 8 show a clear dependence on temporal alignment. Shuffling the visual sequence increases MAE from 2.302 to 2.856 and reduces V-IoU from 0.524 to 0.386. Repeating the first-frame feature removes frame-dependent visual evolution and produces a similar degradation across magnitude, overlap, and temporal consistency. The aligned sequence obtains the best value. These results show that the visual stream contributes time-specific observations that complement the motion representation, rather than acting only as a sequence-level appearance descriptor.

B.5 Robustness to Perturbed Joint Trajectories

We evaluate the sensitivity of KIPS to perturbations in the recovered hand trajectory. At inference time, independent Gaussian noise with standard deviation $\sigma \in \{1, 2, 5\}\,\mathrm{mm}$ is added to each joint coordinate. Velocity and acceleration are then recomputed from the perturbed coordinates, so the experiment evaluates the complete kinematic input rather than perturbing an isolated feature block. The model parameters and visual features remain fixed.

Table 8 reveals different effects on pressure magnitude and contact structure. The pressure MAE remains stable for both variants over the full perturbation range. At $\sigma = 5\,\mathrm{mm}$, the MAE changes by only 0.74% for KIPS-K and 0.36% for KIPS-K+RGB. The spatial overlap metrics, however, show a clear difference between the two variants. For KIPS-K, V-IoU and C-IoU decrease by 9.21% and 9.13%, respectively, at $\sigma = 5\,\mathrm{mm}$. With appearance conditioning, the corresponding changes are limited to 1.45% and 1.08%, while temporal consistency changes by only 1.02%. Thus, temporally aligned appearance provides a complementary observation that stabilizes the predicted contact support when the recovered joint trajectory is perturbed.

B.6 Inference with Limited Lookahead

The main evaluation averages predictions from overlapping temporal windows. We additionally evaluate single-window inference, where each target frame is predicted once without overlap averaging. Lookahead is defined at the input to the pressure estimator: when predicting frame t, a setting with L frames of lookahead allows access to input features up to frame $t + L$. Zero lookahead therefore restricts this input context to the current and preceding frames. We evaluate the original bidirectional model and a unidirectional GRU variant trained with the same pressure prediction

Table 7: Effect of perturbing the temporal organization of the RGB feature stream on EgoTactile.

RGB feature stream	MAE↓	V-IoU↑	C-IoU↑	Temp↑
Aligned	**2.302**	**0.524**	**0.604**	**0.891**
Shuffled	2.856	0.386	0.460	0.810
First frame	3.128	0.192	0.260	0.737
Zeroed	3.124	0.243	0.293	0.804

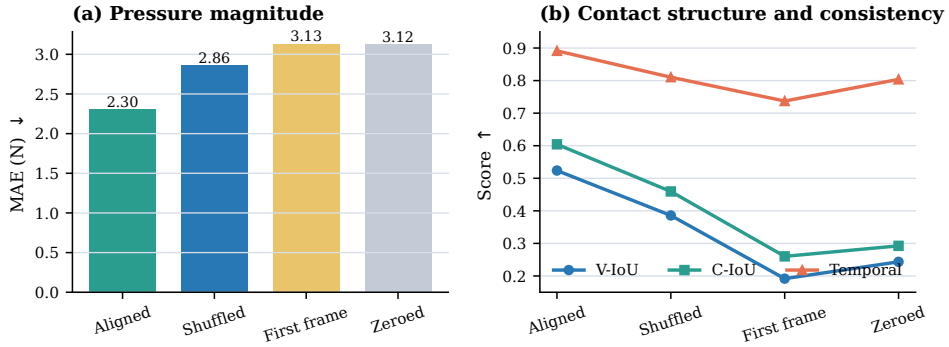


Figure 8: Dependence on temporally aligned visual evidence. **(a)** Pressure MAE. **(b)** Volumetric V-IoU, contact C-IoU, and temporal consistency. Preserving the original alignment gives the strongest pressure and contact estimates.

objective. Figure 10 summarizes the evaluated configurations. Figure 10 reports the low-latency operating points that meet or improve upon the offline MAE.

EgoTactile. With four and eight frames of lookahead, the bidirectional single-window model obtains MAEs of 2.2974 N and 2.2905 N, respectively, compared with 2.3024 N under the offline overlapping-window protocol. The unidirectional GRU variant obtains MAEs of 2.2925 N with two-frame lookahead and 2.2783 N with zero lookahead. Thus, the zero-lookahead configuration achieves a slightly lower MAE than the offline reference without overlap averaging.

OpenTouch. Under the same single-window evaluation protocol, the zero-lookahead configurations obtain MAEs of 1.4912 kPa for KIPS-K and 1.4715 kPa for KIPS-K+RGB. RGB conditioning therefore also improves MAE in this setting.

B.7 Dense Contact Prediction on DexYCB

We additionally evaluate dense hand-contact prediction on the official DexYCB (Chao et al., 2021) S0 split. The evaluation contains $5{,}928$ samples and 778 MANO vertices per sample, resulting in $4{,}611{,}984$ vertex-frame labels. Both methods use the same ground-truth contact labels and a fixed decision threshold of 0.5. Contact vertices account for 25.44% of the evaluated labels.

Table 9 shows that KIPS improves precision from 0.520 to 0.701 and F1 from 0.657 to 0.689. HACO achieves higher recall but predicts contact on 43.78% of the vertices, corresponding to $1.72\times$ the ground-truth contact rate. In comparison, KIPS predicts a contact rate of 24.57%, or $0.97\times$ the ground-truth rate. This more selective prediction reduces false-positive vertex activations by approximately 65% relative to HACO and produces a better overall precision-recall balance.

B.8 Parameter Sensitivity

We examine the sensitivity of KIPS-K to the number of acceleration prototypes and the temporal window length. All experiments are conducted on the EgoTactile gloved object-held-out split with the remaining settings fixed.

Table 8: Robustness to perturbations of the recovered joint coordinates on EgoTactile. Relative changes with respect to the corresponding unperturbed model are shown in parentheses.

Input	σ (mm)	MAE↓	V-IoU↑	C-IoU↑	Temp↑
KIPS-K	0	2.6386	0.4602	0.5464	0.8726
	1	2.6490 (+0.39%)	0.4223 (−8.24%)	0.5026 (−8.02%)	0.8598 (−1.47%)
	2	2.6497 (+0.42%)	0.4211 (−8.50%)	0.5011 (−8.29%)	0.8590 (−1.56%)
	5	2.6581 (+0.74%)	0.4178 (−9.21%)	0.4965 (−9.13%)	0.8561 (−1.89%)
KIPS-K+RGB	0	2.3024	0.5238	0.6043	0.8913
	1	2.3074 (+0.22%)	0.5171 (−1.28%)	0.5982 (−1.01%)	0.8823 (−1.01%)
	2	2.3096 (+0.31%)	0.5166 (−1.37%)	0.5980 (−1.04%)	0.8820 (−1.04%)
	5	2.3106 (+0.36%)	0.5162 (−1.45%)	0.5978 (−1.08%)	0.8822 (−1.02%)

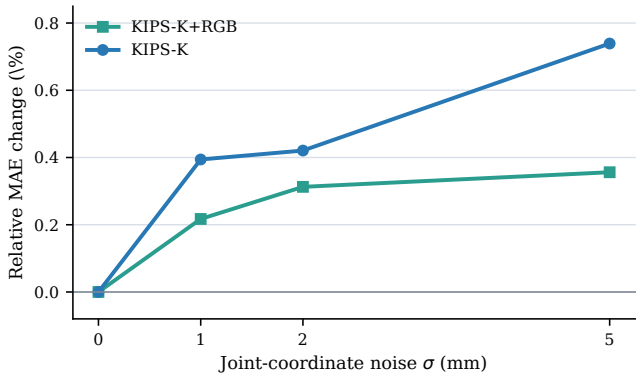


Figure 9: Relative MAE change under increasing joint-coordinate perturbations. Both KIPS-K and KIPS-K+RGB remain stable up to 5 mm noise.

Number of acceleration prototypes. Table 10 shows that $K = 16$ provides the strongest overall balance across pressure magnitude, localization, contact overlap, and temporal consistency. Increasing the dictionary to $K = 32$ reduces MAE by only 0.0212 N, corresponding to a marginal improvement of 0.8%, while producing lower V-IoU, C-IoU, and Temp. Reducing the dictionary to $K = 8$ leads to a clearer degradation across the reported metrics. We therefore use $K = 16$ as the default configuration.

Temporal window length. We evaluate windows of 32, 64, and 96 frames while maintaining a stride equal to half of the window length. The 64-frame configuration achieves the best result on all five metrics. A shorter window provides less temporal context, while extending the window to 96 frames does not provide an additional benefit. These results support the default choice of a 64-frame window with stride 32.

B.9 Geometry of the Learned Acceleration Prototypes

We examine the geometric organization of the learned acceleration dictionary independently of the downstream pressure decoder. For each prototype k, we collect the frames in which it receives the largest joint-averaged assignment and compute the mean gravity direction over those frames. As shown in Figure 11, the learned directions span angles from 19.5° to 122.9°. The dictionary therefore contains motion patterns that are approximately aligned with gravity, nearly orthogonal to gravity, and directed against gravity. This directional coverage shows that the prototype representation does not reduce acceleration to a single magnitude. Instead, the learned keys provide diverse directional anchors that can be softly composed according to the observed joint dynamics.

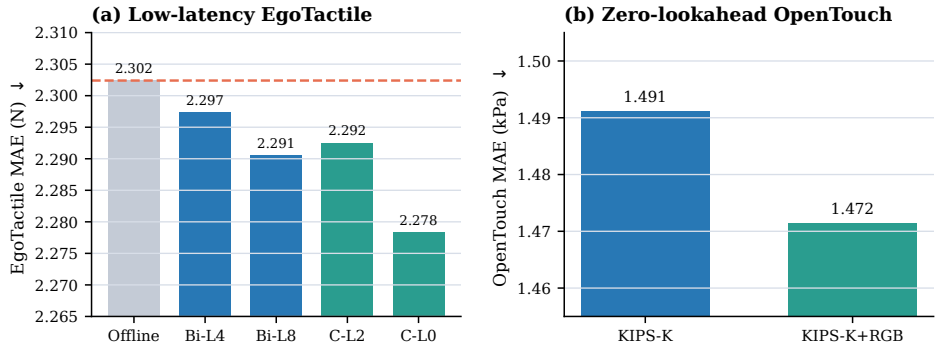


Figure 10: Low-latency pressure inference. **(a)** EgoTactile operating points at or below the standard offline MAE. Bi and C denote bidirectional and causal temporal encoders, and L denotes the number of lookahead frames. **(b)** Zero-lookahead OpenTouch inference with kinematic and appearance-conditioned variants. Both datasets support accurate pressure inference without full future context.

Table 9: Dense contact prediction on the DexYCB S0 split at a fixed threshold of 0.5. The ground-truth contact rate is 25.44%.

Method	Precision↑	Recall↑	F1↑	Pred. contact	Relative to GT
HACO	0.520	**0.894**	0.657	43.78%	1.72×
KIPS	**0.701**	0.677	**0.689**	**24.57%**	**0.97×**

B.10 Efficiency

We further measure the end-to-end efficiency of KIPS-K+RGB on EgoTactile, including HaWoR-based hand recovery, RGB feature extraction, and pressure prediction. The complete pipeline contains 758.78M parameters and requires 292.4 GFLOPs per frame. HaWoR accounts for 670.34M parameters, indicating that the computational footprint is dominated by the upstream hand-recovery model. Under batch-one inference on a single NVIDIA RTX H800 GPU, the complete system processes one frame in 23.11 ms, corresponding to 43.3 FPS. KIPS therefore supports real-time pressure estimation above the 30 FPS input rate even when the full hand-recovery and visual-processing pipeline is included.

B.11 Summary

These experiments extend the main evaluation from pressure accuracy to contact prediction, representation behavior, and practical deployment. KIPS achieves a better precision-recall balance on DexYCB, while the diagnostic analyses support the complementary roles of multi-order kinematics, acceleration prototypes, structured temporal filtering, and aligned visual evidence. Its robustness under trajectory perturbations, causal inference capability, and 43.3 FPS end-to-end throughput further demonstrate its applicability to realistic pressure-estimation settings.

C Limitations

Kinematics provides strong physical evidence for pressure estimation but does not uniquely determine the underlying contact forces. Internal grip strength and unobserved object properties may produce different pressure distributions under similar hand trajectories, which motivates the conditional formulation of KIPS and the use of complementary RGB evidence. The method also depends on the quality of the recovered hand pose, particularly when temporal differentiation amplifies tracking errors. Finally, our evaluation focuses on single-hand manipulation and scalar pressure over the sensor layouts of EgoTactile and OpenTouch. Extending the framework to bimanual interaction, tool use, and directional contact forces requires further study.

Table 10: Parameter sensitivity of KIPS-K on the EgoTactile gloved object-held-out split. The default configuration uses 16 acceleration prototypes and a 64-frame window with stride 32.

Parameter	Setting	MAE↓	CoP↓	V-IoU↑	C-IoU↑	Temp↑
Number of acceleration prototypes						
Prototypes	8	2.7239	3.0446	0.4597	0.5352	0.8618
Prototypes	**16 (default)**	2.6386	**2.9456**	**0.4602**	**0.5464**	**0.8726**
Prototypes	32	**2.6174**	2.9727	0.4565	0.5410	0.8607
Temporal window length and stride						
Window	32 (16)	2.7490	3.0552	0.4496	0.5284	0.8651
Window	**64 (32, default)**	**2.6386**	**2.9456**	**0.4602**	**0.5464**	**0.8726**
Window	96 (48)	2.7862	3.0326	0.4556	0.5300	0.8641

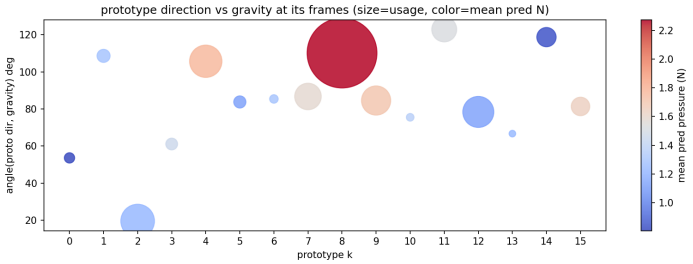


Figure 11: Directional diversity of the learned acceleration prototypes. Each point represents one prototype and shows its angle to the mean gravity direction over the frames in which that prototype is most active. The learned keys cover a broad range of directions relative to gravity.

D Future Work

Future work will investigate the relationship between hand kinematics and contact force at a finer resolution. Promising directions include analyzing how individual joints, motion directions, interaction phases, and temporal delays contribute to pressure at different hand regions. Incorporating object mass, contact normals, and friction may further separate the external force required for object transport from internal grip forces. Controlled experiments that independently vary motion, object load, and grip strength would provide stronger evidence than observational correlation alone. These advances could support physics-aware motion primitives, uncertainty-aware pressure estimation, and the prediction of both normal and tangential forces in more complex interactions.