HOPE: Hand-Object Pressure Estimation from Monocular Videos

Subin Jeon[1] **Byungjun Kim**[1] **Hanbyul Joo**[1,2]
[1]Seoul National University [2]RLWRLD
{subinjeon, byungjun.kim, hbjoo}@snu.ac.kr

https://subin6.github.io/page-hope/

Abstract

Estimating physical pressure from vision is essential for understanding contact-rich hand-object interaction. However, prior vision-based pressure estimation methods are largely limited to planar surfaces and single image input, making them difficult to apply to dynamic hand-object interaction with diverse objects. We instead formulate pressure estimation as a hand-centric video prediction problem with monocular video as input. This formulation predicts temporally evolving per-vertex normal pressure and contact directly on the hand mesh, yielding a unified output space independent of object shape and sensor layout. Building on this formulation, we propose **HOPE**, a framework with two key components. First, we lift tactile-glove pressure, planar-sensor pressure, and distance-based hand-object contact annotations into a shared hand vertex space, allowing bare-hand contact data to regularize pressure learning where labels are unavailable. Second, we introduce a vertex-anchored video transformer that treats each vertex as a persistent token, aggregates visual features and hand pose over time, and uses a contact-gated pressure head to suppress pressure predictions in regions with low contact probability. Experiments on OpenTouch, PressureVisionDB, and hand-object contact benchmarks validate HOPE across object-pressure, surface-pressure, and contact-supervised HOI settings. Despite using pressure supervision primarily from gloved-hand videos, HOPE generalizes to bare-hand egocentric and in-the-wild videos, producing joint contact and pressure predictions beyond the scope of contact-only or planar-pressure baselines.

1 Introduction

Estimating physical pressure from visual observations is a key step toward understanding how humans manipulate the world. While RGB video reveals where hands and objects are, many downstream tasks require knowing how they physically interact: which regions of the hand are in contact, how contact evolves over time, and where pressure is applied during pressing, grasping, or tool use. Such information is important for egocentric perception, augmented and virtual reality, robot learning from human demonstrations, and assistive systems that need to reason about physical interaction beyond object identity or hand pose alone.

Despite this importance, vision-based pressure estimation has so far been studied mostly under constrained sensing geometries. Prior methods [12, 11, 50] have made important progress by estimating pressure from RGB observations in calibrated settings, often using instrumented surfaces where the contact surface is known and pressure can be represented as a 2D planar map. These formulations provide valuable physical supervision, but they remain tied to a particular sensing surface and a limited range of interactions. Everyday hand-object interaction is less structured: the hand contacts arbitrary 3D object surfaces, undergoes frequent self- and object-occlusion, and exhibits

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

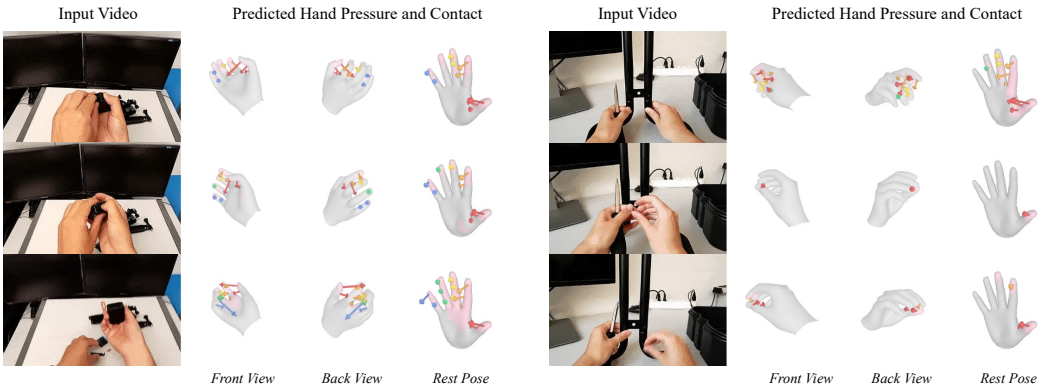


Figure 1: From RGB video to hand-object pressure and contact. **HOPE** predicts vertex-level contact and pressure on the hand mesh without tactile sensing at test time. Pink vertices denote contact, and arrows indicate the averaged pressure computed from the vertex-level pressure predictions.

diverse contact modes such as fingertip pressing, whole-hand grasping, palm support, and tool use. As a result, generalizing pressure estimation beyond instrumented surfaces to arbitrary hand-object interaction remains difficult.

Recent hand-centric tactile datasets for object manipulation, exemplified by OpenTouch [37], mark an important step toward pressure estimation beyond fixed sensing surfaces. By moving tactile sensing from the tactile sensor pad to the hand, they make it possible to collect pressure annotations while the same hand interacts with many different objects. However, this benefit comes with a new limitation: the tactile glove that enables measurement also covers the hand. As a result, many appearance cues used by prior vision-based pressure methods, such as skin deformation, color change, local wrinkles, and fingertip appearance, are no longer directly visible. The visual domain of gloved hands also differs from the bare hands that appear in everyday manipulation, making it difficult for a model trained only on wearable tactile data to generalize to natural bare-hand interaction.

To overcome these limitations, we learn hand-object pressure from complementary pressure and contact supervision. Wearable tactile data provides continuous pressure readings during object manipulation, but is sensor-specific and visually tied to gloved hands; 3D HOI data provides diverse bare-hand interactions, but only contact. We unify these sources by lifting tactile-glove pressure, planar-sensor pressure, and distance-based HOI contact into a shared MANO [34] vertex space, where both pressure and contact are represented as per-vertex fields. Although contact does not determine pressure magnitude, it provides a structural prior on where pressure can occur and supplies bare-hand interaction supervision that gloved pressure data alone cannot provide. This unified representation allows metric pressure labels and broad bare-hand contact data to be used jointly for hand-object pressure estimation.

We further propose a vertex-anchored video transformer for hand pressure estimation. The model represents each MANO vertex as a persistent token, enriches it with visual and hand-pose evidence, and reasons over both the hand surface and time. This design is motivated by the temporal nature of pressure. Pressure is not determined by per-frame appearance alone: it changes as the hand approaches an object, establishes contact, loads the object during grasping or holding, and releases it. In these phases, visually similar contact configurations may correspond to different pressure magnitudes. Temporal modeling therefore helps disambiguate loading states and propagates pressure cues through short occlusions. Finally, a contact-gated pressure head couples the two predictions by suppressing pressure in regions with low predicted contact probability.

We evaluate the proposed framework on hand-object pressure estimation with OpenTouch [37], hand-surface pressure estimation with PressureVisionDB [12], and dense hand contact estimation on hand-object interaction [10, 5]. Our experiments show that the proposed formulation enables pressure estimation beyond fixed planar surfaces and benefits from combining pressure supervision with contact-only HOI data. Furthermore, as illustrated in Fig. 1, the learned hand-centric representation can be applied to bare-hand videos, producing dense contact and pressure predictions despite pressure supervision being obtained primarily from gloved-hand videos.

Our contributions are threefold. First, we introduce a hand-centric formulation for visual hand-object pressure estimation, representing pressure and contact as per-vertex fields on the MANO hand mesh. Second, we unify heterogeneous supervision by lifting tactile-glove pressure, planar-sensor pressure, and contact-only HOI annotations into the same vertex space, allowing bare-hand contact data to serve as a structural prior for pressure estimation. Third, we propose a vertex-anchored video transformer with contact-gated pressure prediction, enabling temporally informed pressure estimation from monocular RGB video.

2 Related Work

2.1 Vision-based hand pressure estimation.

Estimating physical pressure from visual observations has recently emerged as a way to perceive contact-rich interaction without instrumenting every object or surface. PressureVision and its follow-up showed that RGB images contain pressure cues such as skin deformation, color change, pose, and shadows, while EgoPressure moves this problem into the egocentric setting with paired pressure and pose annotations [12, 11, 50]. These works establish that pressure can be inferred from visual evidence, but their pressure outputs remain tied to planar sensor images, image-space maps, fingertip-level labels, or a particular capture setup. Among them, EgoPressure [50] is the most closely related to ours: both build on off-the-shelf MANO reconstruction to express pressure in a hand-centric space. EgoPressure, however, estimates bare-hand pressure frame-wise against an instrumented planar surface, whereas HOPE targets diverse hand-object interactions, jointly learns from heterogeneous pressure and contact supervision, and models temporal pressure evolution.

2.2 Tactile sensing and touch datasets for hand-object interaction.

Direct tactile sensing provides physical supervision that is difficult to obtain from vision alone. Physical touch has been captured with instrumented objects or sensing surfaces, wearable tactile gloves, and thermal contact imprints [31, 12, 38, 3, 4]. Such setups provide direct force, pressure, or contact evidence, but their labels are naturally expressed in the coordinate system and coverage of the underlying hardware, such as force transducer locations, planar sensels, thermal object surfaces, or glove taxels. In contrast, many large hand-object interaction datasets provide hand-object pose, geometry, 2D contact, or inferred contact annotations that support contact reasoning, but do not directly measure pressure [17, 35, 5, 39, 7, 14, 15, 24, 28, 43, 10, 2]. Most closely related to our setting, OpenTouch [37] captures in-the-wild egocentric video, full-hand tactile maps, and hand pose with a wearable tactile glove, making full-hand touch available during natural manipulation. We use these physical signals as complementary supervision, but lift them into a shared MANO vertex representation so that heterogeneous pressure and contact labels can be learned jointly.

2.3 Hand-object contact and affordance estimation.

Hand-object interaction has also been widely studied through contact, grasp, and affordance rather than pressure. Contact cues have been used to annotate grasp regions, refine hand-object poses, generate plausible grasps, and reconstruct in-the-wild hand-object interactions [3, 4, 20, 13, 26, 18, 27, 30]. Large-scale interaction datasets and contact-estimation pipelines further show that contact can be derived from tracked geometry across diverse hand-object settings, although these labels are typically binary or proximity-based rather than metric pressure [17, 35, 5, 39, 7, 14, 15, 24, 28, 43, 10]. Dense vertex-level contact has also been studied for human-scene interaction, while HACO focuses on MANO-vertex hand contact from RGB images under severe class and spatial imbalance [16, 19, 41, 21]. Beyond static contact labels, HOI generation and affordance models synthesize hand-object motion, contact maps, or object-centric interaction regions from text, visual observations, object geometry, and learned interaction priors [8, 6, 25, 40, 42, 46, 47, 45, 48, 1, 44, 33, 23, 22]. Among them, LatentAct is particularly close to our representation because it predicts future 3D hand motion and MANO contact-map trajectories, but it still models where contact occurs rather than how much pressure is applied [33]. These contact and affordance models provide important cues about where interaction is likely to occur, but they do not estimate the magnitude of physical loading on the hand. Our work treats contact as a structural prior for pressure and jointly predicts contact and per-vertex pressure, thereby connecting contact-rich HOI understanding with hand pressure estimation.

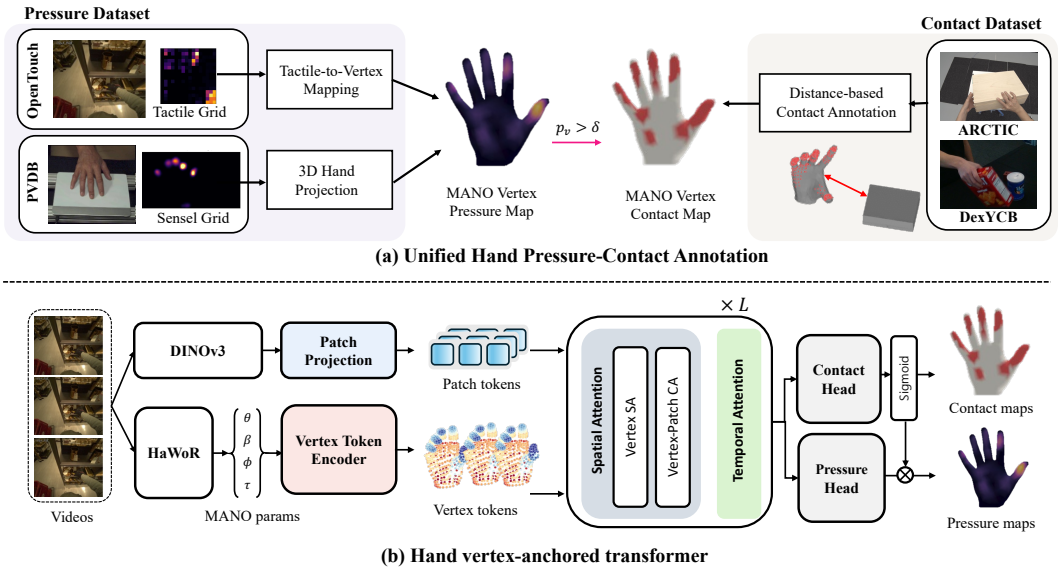


Figure 2: **Method overview.** (a) **Unified Hand Pressure–Contact Annotation.** Heterogeneous pressure and contact supervision is lifted into a shared MANO vertex space. (b) **VertexFormer.** Image patch tokens and vertex tokens are fused via interleaved spatial and temporal attention.

3 Method

We propose **HOPE**, a framework for predicting per-vertex contact and pressure on the hand surface from RGB video. Given a clip of T frames $\{I_t\}_{t=1}^{T}$, HOPE predicts per-vertex pressure $\hat{\mathbf{p}}_t \in \mathbb{R}^V_{\geq 0}$ and per-vertex contact $\hat{\mathbf{c}}_t \in [0,1]^V$ on the MANO [34] mesh, where V=778. HOPE consists of two components: (i) a unified annotation pipeline that lifts tactile-glove pressure, planar-sensor pressure, and distance-based hand-object contact into a shared MANO vertex space (Sec. 3.1), and (ii) **VertexFormer**, a vertex-anchored video transformer that predicts contact and pressure jointly on the MANO mesh through interleaved spatial and temporal attention with a contact-gated pressure head (Sec. 3.2). Training is performed with a unified objective: samples with pressure labels supervise both per-vertex pressure regression and contact classification, while contact-only HOI samples supervise only the contact head (Sec. 3.3).

3.1 Unified pressure–contact annotation on MANO vertex space

Existing pressure datasets provide annotations in sensor-specific coordinate systems—OPENTOUCH [37] in a 16×16 tactile taxel grid attached to the palmar side of a glove, and PressureVisionDB (PVDB) [12] in a 2D sensor image of a planar touchpad. Training a single model across both sources requires a common output space. Moreover, while these datasets provide direct pressure supervision, their sensing setups cover constrained interaction domains. We therefore complement them with distance-based contact annotations from existing HOI datasets [10, 21], which provide diverse bare-hand object interactions without pressure labels (Fig. 2 (a)).

Tactile-to-vertex mapping (OPENTOUCH). OPENTOUCH provides a correspondence $\mathcal{M} : 1, \ldots, 256 \to 2^{\mathcal{V}}$ between each taxel and the MANO vertices physically covered by its sensing patch, which we project onto the standard V=778 topology. Since taxel patches span many vertices (38 on average) and adjacent patches share vertices at their boundaries, we aggregate the taxel pressures $\boldsymbol{\pi} \in \mathbb{R}^{256}_{\geq 0}$ at each vertex by averaging over its covering taxels,

$$\tilde{p}_v = \frac{1}{|\mathcal{T}(v)|} \sum_{i \in \mathcal{T}(v)} \pi_i, \qquad \mathcal{T}(v) = \{i : v \in \mathcal{M}(i)\}, \tag{1}$$

yielding a vertex-level pressure field $\mathbf{p}^{\text{OT}} \in \mathbb{R}^V_{\geq 0}$ over the sensor-covered region of the hand.

Sensel-to-vertex projection (PVDB). PVDB provides dense pressure on a planar sensor pad synchronized with RGB video, but does not include 3D hand annotations. We obtain a posed MANO mesh per frame from an off-the-shelf hand reconstruction model, refined with multi-view fusion. We then project each palmar vertex onto the pad plane using the calibrated camera-to-pad extrinsics and bilinearly sample the sensor pressure map at its (u, v) coordinate, yielding $\mathbf{p}^{\text{PV}} \in \mathbb{R}^{V}_{\geq 0}$.

Pressure and contact supervision. PVDB provides pressure directly in kPa, while OPENTOUCH provides tactile readings that we linearly convert to kPa using its reported pressure range. We use the resulting vertex-level values as pressure supervision. Additionally, we derive binary contact labels by thresholding the pressure, $c_{t,v} = \mathbf{1}[p_{t,v} > \tau_p]$.

Distance-based contact from HOI data. To complement pressure supervision with diverse bare-hand object interactions, we incorporate contact annotations from existing HOI datasets (DexYCB [7], ARCTIC [10]) following the protocol of HACO [21]. Given the ground-truth posed hand and object meshes, a vertex is labeled as in contact if its distance to the object surface is below a threshold δ_c,

$$c_{t,v} = \mathbf{1}\Big[\operatorname{dist}\Big(\mathbf{x}^{\text{hand}}_{t,v},\, \mathcal{S}^{\text{obj}}_{t}\Big) < \delta_c\Big], \tag{2}$$

where $\mathbf{x}^{\text{hand}}_{t,v}$ is the world-space coordinate of vertex v at frame t and $\mathcal{S}^{\text{obj}}_{t}$ is the object surface. These annotations provide complementary contact supervision without requiring pressure measurements. The final training set is the union of pressure samples (with p_t and derived c_t) from OpenTouch/PVDB and contact-only samples (with c_t alone) from HOI datasets, all expressed in the same V-dimensional MANO vertex space.

3.2 VertexFormer: a vertex-anchored transformer

VertexFormer (Fig. 2 (b)) tokenizes the input along two complementary axes—image patches and hand vertices—and reasons over both with interleaved spatial and temporal attention. Throughout the network, vertex tokens carry the prediction state and image patches serve as a key–value memory; we refer to this as a *vertex-anchored* design.

Patch and vertex tokens. Each frame is encoded by a frozen ViT image encoder [36] into a grid of patch features, which are linearly projected to dimension d to form patch tokens $\mathbf{F}t \in \mathbb{R}^{N_p \times d}$. In parallel, an off-the-shelf 3D hand reconstruction model estimates per-frame MANO parameters $(\theta_t, \beta_t, \phi_t, \tau_t)$, where θ_t and β_t denote pose and shape, and ϕ_t and τ_t denote global orientation and camera-relative translation, respectively. In our implementation, we use the camera-space MANO estimates from HaWoR [49], prior to its world-space reconstruction. To support fine-grained pressure prediction directly on the hand surface, we embed the posed MANO vertices as persistent vertex tokens. We first obtain hand-local vertex coordinates $\tilde{\mathbf{V}}t \in \mathbb{R}^{V\times 3}$ from the estimated pose and shape, and construct each vertex token as

$$\mathbf{z}_{t,v} = \mathbf{e}_v + W_v\, \tilde{\mathbf{V}}_{t,v} + W_\phi\, \phi_t + W_\tau\, \tau_t, \tag{3}$$

where $\mathbf{e}_v$ is a learned vertex-identity embedding, and W_v, W_ϕ, and W_τ are learned projections of the hand-local vertex coordinate, global orientation, and translation, respectively, into the token dimension d. The identity embedding maintains a fixed correspondence between tokens and MANO vertices, while separating hand-local geometry from global orientation and translation allows the model to reason separately about hand articulation, viewpoint, and camera-relative position.

Spatial vertex–patch attention. Each transformer block applies two attention operations over the vertex tokens. *Vertex self-attention* (Vertex-SA) lets vertices exchange information along the hand topology, propagating contact signals from observable regions to occluded ones. *Vertex-to-patch cross-attention* (Vertex-Patch CA) lets each vertex token attend to the full patch grid, aggregating visual evidence from the image. Together these realize the vertex-anchored design: vertices are queries and remain the carriers of state, while image features serve as a key–value memory.

Temporal attention. A subsequent temporal attention layer attends over the T frames per vertex index, sharing information across time at fixed vertex identity. This aggregates temporal evidence for pressure prediction and helps disambiguate contact during brief occlusions. The spatial–temporal block is repeated L times.

Contact-gated pressure prediction. Two MLP heads take the final vertex tokens as input. The contact head outputs a logit per vertex, passed through a sigmoid to produce $\hat{\mathbf{c}}_t$. The pressure head

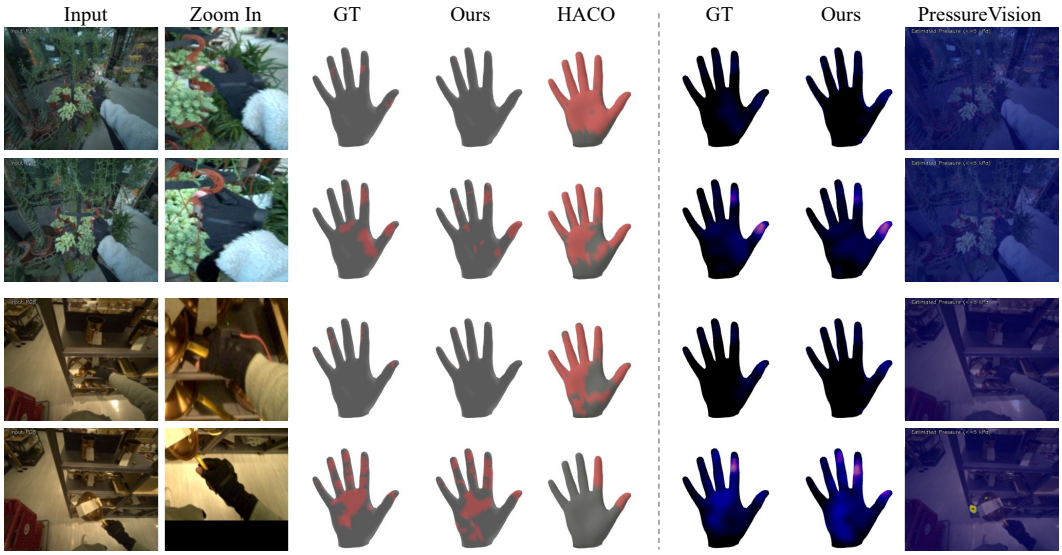


Figure 3: **Qualitative comparison on OpenTouch.** (a) Vertex-level contact (b) Vertex-level pressure

Table 1: Quantitative comparison of hand-object pressure estimation on OpenTouch dataset.

	Contact (frame-level)			Contact (vertex-level)			Pressure (vertex-level)	
	F1	Precision	Recall	F1	Precision	Recall	MAE (kPa)	RMSE (kPa)
PressureVision [12]	0.164	0.767	0.092	0.013	0.239	0.007	1.55 (5.96 / 0.18)	4.43 (9.05 / 0.59)
PressureVision++ [11]	0.008	0.923	0.004	0.001	0.404	0.001	1.53 (5.98 / 0.16)	4.44 (9.07 / 0.59)
HACO [21]	0.801	0.669	1.000	0.373	0.267	0.615	-	-
Ours	0.888	0.824	0.964	0.680	0.691	0.669	1.23 (3.73 / 0.45)	3.10 (5.87 / 1.38)

outputs a non-negative pressure value per vertex, $\tilde{\mathbf{p}}_t$. We then gate the pressure prediction by the contact probability,

$$\hat{\mathbf{p}}_t \;=\; \hat{\mathbf{c}}_t \odot \tilde{\mathbf{p}}_t, \tag{4}$$

where $\odot$ denotes element-wise product over vertices. The gate suppresses pressure where the predicted contact probability is low, allowing contact-only HOI supervision to guide pressure localization.

3.3 Training objective

Let $\Omega_t^p \subseteq \{1, \ldots, V\}$ and Ω_t^c denote the sets of vertices with valid pressure and contact annotations, respectively. We normalize pressure using a shared reference scale $s = 50\,\text{kPa}$ (the nominal full-scale range of OpenTouch) for both OPENTOUCH and PVDB, defining $q_{t,v} = p_{t,v}/s$ and $\hat{q}_{t,v} = \hat{p}_{t,v}/s$. We supervise contact with binary cross-entropy and pressure with mean squared error (MSE):

$$\mathcal{L}_{\text{contact}} = \frac{1}{|\Omega_t^c|} \sum_{v \in \Omega_t^c} \text{BCE}(\hat{c}_{t,v}, c_{t,v}), \qquad \mathcal{L}_{\text{pressure}} = \frac{1}{|\Omega_t^p|} \sum_{v \in \Omega_t^p} (\hat{q}_{t,v} - q_{t,v})^2. \tag{5}$$

The total objective is $\mathcal{L} = \lambda_c \mathcal{L}_{\text{contact}} + \lambda_p \mathcal{L}_{\text{pressure}}$, where the pressure term is omitted for contact-only HOI samples. This unified loss enables joint training over pressure and contact data despite their heterogeneous supervision coverage.

4 Experiments

4.1 Experimental setup

Datasets. We train and evaluate our model on a combination of pressure and contact datasets that cover different forms of hand interaction. For hand-object pressure, we use OPENTOUCH [37], which contains egocentric videos of natural hand-object interaction with synchronized tactile measurements

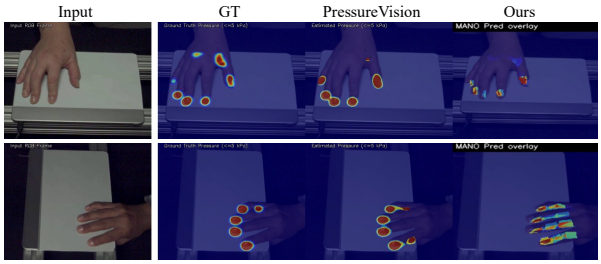


Figure 4: Qualitative comparison on PressureVisionDB. Vertex-level predictions are projected to the image plane.

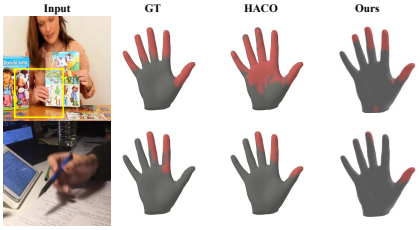


Figure 5: Contact comparison on MOW dataset.

Table 2: Hand-surface pressure estimation on PressureVisionDB. MAE/RMSE are vertex-level errors in kPa; IoUs are pixel-level.

	Contact		Pressure		
	Frame F1 (P / R)	contact IoU	vol IoU	MAE	RMSE
PV [12]	0.939 (0.977 / 0.904)	0.547	0.411	0.471	3.726
PV++ [11]	0.248 (0.963 / 0.143)	0.022	0.021	0.248	3.025
Ours	0.903 (0.901 / 0.906)	0.326	0.210	0.303	2.073

Table 3: Dense hand-object contact estimation on MOW dataset.

	Precision	Recall	F1
POSA [16]	0.134	0.128	0.101
BSTRO [19]	0.204	0.126	0.112
DECO [41]	0.246	0.235	0.197
HACO [21]	0.525	0.607	0.522
Ours	0.549	0.420	0.422

from a wearable glove and hand pose annotations. We use 1,644 clips / 171,765 frames for training and 190 clips / 17,264 frames for testing. For hand-surface pressure, we use PressureVisionDB (PVDB) [12], which contains RGB observations synchronized with pressure maps measured on a planar sensor surface; we use 815 sequences / 170,720 frames for training and 204 sequences / 33,280 frames for testing. For contact supervision, we additionally use DEXYCB [7] (1,144 clips / 10,786 frames) and ARCTIC [10] (193 clips / 74,705 frames) with distance-based contact labels. MOW [5] (92 images, evaluation only), which provides single-image hand-object interaction in the wild, is used for image-level contact evaluation following the protocol of HACO [21].

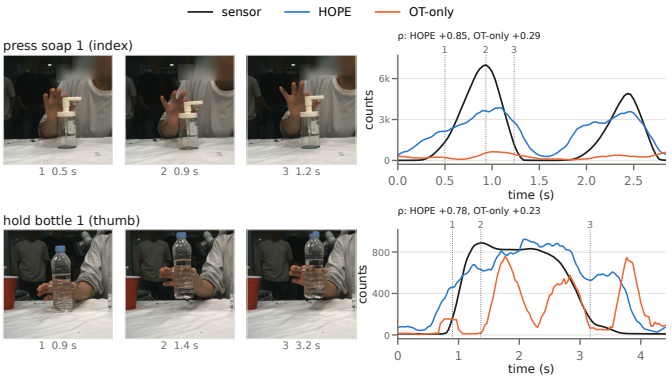
Baselines. We compare with prior methods according to their output space and task. For hand-surface pressure estimation, we compare against PressureVision [12] and PressureVision++ [11], which predict pressure on planar image coordinates. For dense hand contact estimation, we compare against HACO [21], which predicts MANO vertex-level hand contact from RGB images. Since HACO does not estimate pressure magnitude, we report only contact metrics for HACO. For all baselines, we use the official checkpoints released by the authors without further fine-tuning.

Metrics. We report contact and pressure metrics at multiple granularities. *Frame-level contact* F1/Precision/Recall treats each frame as a binary classification (contact present anywhere on the hand vs. not). *Vertex-level contact* F1/P/R is computed per-vertex across all frames, measuring spatial localization of contact. For pressure, we report *vertex-level mean absolute error (MAE) and root mean square error (RMSE)* in kPa, computed over vertices with valid pressure annotations (Ω_t^p; glove-covered vertices for OpenTouch and all vertices for PVDB). For OpenTouch, we additionally decompose MAE/RMSE into contact / non-contact regions (numbers in parentheses) to separately assess pressure magnitude estimation and false-positive suppression. For PVDB, we additionally report *pixel-level contact IoU* and *volume IoU* on the pad plane by projecting MANO-vertex predictions back to sensor coordinates, following the PressureVision protocol.

Implementation details. We use DINOv3 [36] ViT-B/16 [9] as a visual backbone and HaWoR [49] for per-frame MANO parameter estimation; both are kept frozen during training. Input frames are resized to 224×224, yielding $N_p = 196$ patch tokens per frame. Each clip contains $T = 10$ frames. The vertex-anchored transformer uses $L = 2$ spatio-temporal blocks with hidden dimension $d = 64$ and 4 attention heads. We train with AdamW [29] with a cosine learning-rate schedule, learning rate $3{\times}10^{-3}$, batch size 128, for 20 epochs on $1{\times}$RTX 3090. Loss weights are set to $\lambda_c = 1.0$ and $\lambda_p = 1.0$; the pressure loss is masked out for contact-only HOI samples. The contact threshold is $\tau_p = 1$ kPa and the distance-based contact threshold follows HACO at $\delta_c = 10$ mm.

4.2 Comparison

Hand-object pressure on OpenTouch. Table 1 reports results on the held-out OpenTouch test set, our main target setting. Our method achieves the highest vertex-level contact F1 and the lowest pressure



(a) Temporal pressure predictions

	Combined	Thumb	Index
Mean Spearman's ρ			
PV	0.13	0.00	−0.06
PV++	0.00	−0.06	−0.17
OT-only	0.17	0.21	0.02
HOPE	**0.50**	**0.42**	**0.35**
Positive clips ($\rho > 0$)			
PV	14/23	12/23	8/23
PV++	13/23	10/23	9/23
OT-only	15/23	17/23	12/23
HOPE	**23/23**	**22/23**	**21/23**

(b) Quantitative evaluation

Figure 6: Near-bare-hand pressure evaluation. (a) Selected RGB frames and corresponding fingertip sensor readings (black), compared with HOPE (blue) and OpenTouch-only (orange). (b) Mean clip-wise Spearman correlation and number of clips with positive correlation. For visualization, predictions are rescaled to sensor counts per clip and model. ρ is computed before rescaling.

Table 4: Dataset composition ablation. **OT**: OpenTouch (pressure), **PV**: PressureVisionDB (pressure), **HOI**: DexYCB + ARCTIC (contact). We evaluate on OpenTouch (in-domain hand-object pressure), PVDB (hand-surface pressure), and ARCTIC (bare-hand contact generalization).

Training data			OpenTouch			PVDB		ARCTIC	
OT	PV	HOI	Frame F1 ↑	Vertex F1 ↑	MAE ↓	Frame F1 ↑	MAE ↓	Frame F1 ↑	Vertex F1 ↑
✓			0.886	0.669	1.239	0.729	1.094	0.825	0.157
✓	✓		0.886	**0.687**	1.231	**0.910**	**0.257**	0.737	0.087
✓	✓	✓	**0.888**	0.680	**1.226**	0.903	0.303	**0.954**	**0.513**

errors among methods that predict pressure. PressureVision predicts pressure in a 2D planar map and cannot transfer meaningfully to MANO vertices, which explains its near-zero vertex-level F1. HACO predicts MANO-vertex contact and achieves high recall, but its precision drops substantially when applied to gloved-hand interaction, indicating a tendency to over-predict contact regions. Qualitative results in Fig. 3 confirm that our predictions concentrate pressure on physically meaningful contact regions and align spatially with the ground-truth tactile map.

Hand-surface pressure on PVDB. We next evaluate on PressureVisionDB to assess whether the same MANO-vertex representation supports planar hand-surface pressure estimation (Table 2). Our method is competitive with PressureVision on frame-level contact and achieves the lowest pressure RMSE, despite predicting in a hand-centric space rather than the sensor's native 2D image space. PressureVision retains an advantage on pixel-level contact IoU, as projecting MANO-vertex predictions back to the pad plane introduces hand reconstruction error that planar methods avoid. Overall, these results suggest that the MANO-vertex formulation supports planar-surface pressure estimation while enabling pressure estimation in non-planar settings. Qualitative comparisons are shown in Fig. 4.

Dense hand-object contact on MOW. We also evaluate contact transfer to in-the-wild bare-hand object interaction on MOW (Table 3). Our model achieves higher precision than HACO but lower recall, resulting in a lower overall F1. Qualitative results show that HOPE tends to predict more localized contact regions, consistent with the difference between pressure-derived and proximity-based contact supervision (Fig. 5). Importantly, our model maintains reasonable contact performance while additionally enabling pressure-magnitude prediction, a capability absent in contact-only baselines.

Pressure generalization to near-bare-hand interactions. To quantitatively assess HOPE's generalization beyond gloved-hand training data, we collect 23 held-out clips (6,310 frames) of everyday object manipulation. Two pressure sensors are attached to the thumb and index fingertips, leaving most of the hand visually uncovered. Since the sensors provide uncalibrated readings, we measure temporal agreement using clip-wise Spearman's rank correlation (ρ). We compare HOPE against PressureVision, PressureVision++, and an *OpenTouch-only* variant trained without the additional heterogeneous supervision. As shown in Fig. 6, HOPE achieves the highest mean correlation across

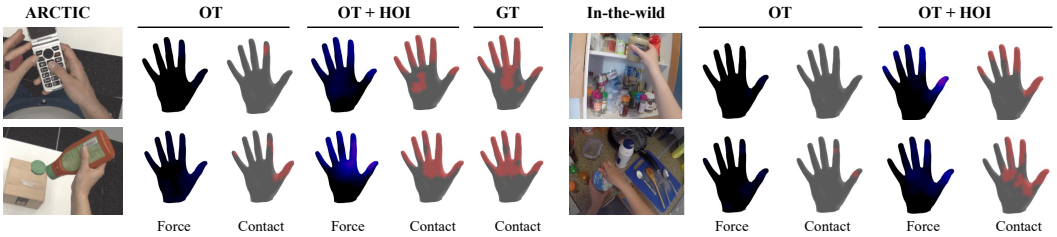


Figure 7: Generalization to bare-hand object interaction. We compare predictions from models trained without (OT) and with (OT + HOI) bare-hand contact supervision.

Table 5: Ablation study on input modalities. We compare variants using different combinations of visual (RGB), hand pose, and temporal information. **Bold** indicates the best result.

Variant	Inputs			Frame F1	Vertex-level			MAE (kPa)
	Visual	Pose	Temp.	↑	F1 ↑	P ↑	R ↑	↓
Pose only	×	✓	×	0.8337	0.6161	0.6341	0.5991	1.4472
RGB only	✓	×	×	0.8567	0.6363	0.6480	0.6250	1.3601
RGB + Pose	✓	✓	×	0.8656	0.6566	0.6839	0.6314	1.2947
Pose sequence	×	✓	✓	0.8501	0.6132	0.6756	0.5614	1.3969
Video	✓	×	✓	0.8822	0.6658	0.6758	0.6561	1.2873
Full (Ours)	✓	✓	✓	**0.8884**	**0.6795**	**0.6905**	**0.6689**	**1.2261**

the overall measure and both fingertips, with positive correlations on nearly all clips. Compared with OpenTouch-only, HOPE improves the overall correlation from 0.17 to 0.50, outperforming it on 22 of the 23 clips (paired Wilcoxon signed-rank test, $p = 2.5 \times 10^{-5}$). The temporal examples further illustrate that HOPE more closely follows the measured pressure buildup and release during object manipulation. These results suggest that heterogeneous supervision improves the transfer of relative pressure dynamics to near-bare-hand interactions.

4.3 Analysis

Effect of training data composition. We study how each training source contributes to in-domain performance and cross-domain generalization (Table 4). Training on OpenTouch alone (OT) already provides strong hand-object pressure performance on the OpenTouch test set, but transfers poorly to the planar PVDB setting and gives only modest contact accuracy on bare-hand ARCTIC interactions. Adding PVDB pressure supervision (OT + PV) substantially improves planar hand-surface pressure estimation while leaving in-domain OpenTouch performance largely unchanged, indicating that the unified MANO-vertex representation absorbs heterogeneous pressure sources without interference. Since both OT and PV expose the model only to gloved or planar-surface interactions, contact transfer to ARCTIC remains limited. Incorporating distance-based contact from HOI datasets (OT + PV + HOI) substantially improves bare-hand contact localization on ARCTIC, while slightly increasing pressure error on PVDB. These results highlight the complementary role of contact-only supervision in extending pressure estimation to diverse bare-hand interactions. Qualitative comparisons in Fig. 7 show the corresponding effect on bare-hand transfer.

Input modality ablation. We ablate the three input modalities of the vertex-anchored transformer—visual features, hand pose, and temporal context—in Table 5: RGB is removed by skipping the vertex-to-patch cross-attention, pose by replacing the posed MANO vertices with the canonical-pose template (leaving only the per-vertex identity embedding e_v), and temporal context by disabling the temporal attention layer. Pose-only and RGB-only variants achieve comparable performance, and combining the two consistently improves over either alone, indicating that hand articulation and visual evidence carry complementary, non-redundant signals. Temporal context further improves the RGB-based variants, supporting the importance of modeling pressure evolution over time. The full model achieves the best performance across the reported contact and pressure metrics.

Vertex token ablation. To assess the effectiveness of vertex-level reasoning, we compare our vertex-token representation with a compact alternative using 18 MANO parameter tokens instead of 778 vertex tokens. In this variant, spatial and temporal attention operate on parameter tokens, whose features are subsequently expanded to 778 vertex features using the fixed linear blend skinning (LBS) weights of MANO. As shown in Table 6(a), vertex tokens improve contact F1 from 0.634 to

Table 6: Architecture ablations. (a) Comparison of vertex-token and parameter-token representations. (b) Retraining with different hand reconstructors and visual backbones.

Representation	Vertex F1 ↑	MAE ↓
Parameter tokens + LBS	0.634	1.237
Vertex tokens (Ours)	**0.680**	**1.226**

(a) Token representation

Reconstructor + backbone	Vertex F1 ↑	MAE ↓
HaWoR + DINOv3 (default)	0.680	1.226
WiLoR + DINOv3	0.661	1.280
HaWoR + HaWoR ViT-H	0.679	1.178

(b) Component replacement

Table 7: Robustness to test-time MANO perturbations on OpenTouch. Each perturbation is drawn once per clip and applied to every frame. The default model reaches 0.680 vertex F1 and 1.226 kPa MAE. We report the change (Δ) from the default.

	Global trans. (mm)				Global orient. (°)				Joint pose (°)				Shape (std)		
	10	20	50	100	5	10	20	40	2	5	10	20	0.5	1	2
Δ Vertex F1 ↑	+0.0001	+0.0002	+0.0003	−0.0004	0.0000	−0.0001	−0.0006	−0.0040	−0.0007	−0.0024	−0.0065	−0.0172	−0.0010	−0.0019	−0.0046
Δ MAE (kPa) ↓	−0.0001	−0.0002	−0.0005	−0.0009	−0.0001	+0.0006	+0.0043	+0.0195	+0.0013	+0.0047	+0.0131	+0.0336	+0.0004	+0.0014	+0.0049

0.680, with slightly lower pressure MAE. These results support our vertex-anchored design, where vertex-specific representations are maintained throughout the network for fine-grained reasoning over the hand surface.

Robustness to hand reconstruction. Since HOPE relies on MANO estimates from an off-the-shelf reconstructor, we assess its sensitivity to reconstruction errors (Table 7). Injecting random noise into the estimated MANO parameters at test time causes limited degradation under moderate perturbations, with the largest F1 drop of 0.017 under 20° per-joint pose noise. The model is trained on estimated MANO parameters, so it already sees reconstruction error during training and learns to tolerate it. Replacing HaWoR with the single-image reconstructor WiLoR [32] slightly reduces contact F1 (0.680 to 0.661) and increases pressure MAE (1.226 to 1.280 kPa), suggesting that reliable hand reconstruction can benefit downstream pressure estimation. Separately, replacing DINOv3 with HaWoR's ViT-H features achieves comparable contact F1 (0.679) and lower pressure MAE (1.178 kPa), indicating that VertexFormer is not tied to a particular visual backbone (Table 6(b)).

5 Discussion and conclusion

We presented HOPE, a framework for predicting per-vertex normal pressure and contact from monocular videos. By lifting tactile-glove pressure, planar-sensor pressure, and distance-based contact annotations into a shared MANO vertex space, HOPE enables joint learning from heterogeneous supervision. Our experiments show that contact-only HOI supervision improves contact localization in bare-hand interactions, while the near-bare-hand evaluation further demonstrates improved agreement with measured temporal pressure variations. These results support the effectiveness of combining complementary pressure and contact supervision for hand-object pressure estimation beyond instrumented settings.

Limitations. HOPE relies on an off-the-shelf hand reconstructor and may inherit its errors under severe occlusion or challenging viewpoints, although our experiments show robustness to moderate reconstruction noise. Our training supervision also has limitations in sensing coverage and label consistency: tactile measurements cover only instrumented regions, while distance-based contact annotations do not necessarily correspond to physically loaded contact. Furthermore, our near-bare-hand evaluation uses two uncalibrated fingertip sensors, allowing us to assess relative temporal pressure variations but not absolute pressure accuracy in kPa. Finally, HOPE predicts only normal pressure, leaving tangential forces and full contact wrench estimation for future work.

Acknowledgments and Disclosure of Funding

This work was supported by a research grant from the Youlchon Foundation, RLWRLD, and Institute of Engineering Research at Seoul National University, NRF grant funded by the Korean government (MSIT) [No. RS-2022-NR070498 and No. RS-2026-25540546], and IITP grant funded by the Korea government (MSIT) [No. RS-2024-00439854, No. RS-2025-25441838, No. RS-2025-25442338, and No. RS-2021-II211343]. H. Joo is the corresponding author.

References

[1] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[2] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[3] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. ContactDB: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[4] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[5] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[6] Junuk Cha, Jihyeon Kim, Jae-Shin Yoon, and Seungryul Baek. Text2HOI: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[7] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[8] Sammy Joe Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. DiffH2O: Diffusion-based synthesis of hand-object interactions from textual descriptions. In *SIGGRAPH Asia 2024 Conference Papers*, 2024.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[10] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[11] 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[12] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[13] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[14] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint Transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[16] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3d scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[17] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[18] Junxing Hu, Hongwen Zhang, Zerui Chen, Mengcheng Li, Yunlong Wang, Yebin Liu, and Zhenan Sun. Learning explicit contact for implicit reconstruction of hand-held objects from monocular images. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

[19] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[20] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

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

[22] Hyeonwoo Kim, Sangwon Baik, and Hanbyul Joo. DAViD: Modeling dynamic affordance of 3d objects using pre-trained video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[23] Hyeonwoo Kim, Sookwan Han, Patrick Kwon, and Hanbyul Joo. Beyond the contact: Discovering comprehensive affordance for 3d objects from pre-trained 2d diffusion models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[24] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2O: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[25] Quanzhou Li, Jingbo Wang, Chen Change Loy, and Bo Dai. Task-oriented human-object interactions generation with implicit neural representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[26] Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. ContactGen: Generative contact modeling for grasp generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[27] Yumeng Liu, Xiaoxiao Long, Zemin Yang, Yuan Liu, Marc Habermann, Christian Theobalt, Yuexin Ma, and Wenping Wang. EasyHOI: Unleashing the power of large models for reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[28] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[30] Hyeongjin Nam, Daniel Sungho Jung, Gyeongsik Moon, and Kyoung Mu Lee. Joint reconstruction of 3d human and object via contact-based refinement transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[31] 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 (TPAMI)*, 40(12):2883–2896, 2018.

[32] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12242–12254. IEEE, 2025.

[33] Aditya Prakash, Benjamin Lundell, Dmitry Andreychuk, David Forsyth, Saurabh Gupta, and Harpreet Sawhney. How do i do that? synthesizing 3d hand motion and contacts for everyday interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[34] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):245:1–245:17, 2017.

[35] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F. Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[36] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[37] Yuxin Ray Song, Jinzhou Li, Rao Fu, Devin Murphy, Kaichen Zhou, Rishi Shiv, Yaqi Li, Haoyu Xiong, Crystal Elaine Owens, Yilun Du, Yiyue Luo, Xianyi Cheng, Antonio Torralba, Wojciech Matusik, and Paul Pu Liang. Opentouch: Bringing full-hand touch to real-world interaction, 2025.

[38] 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.

[39] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[40] Omid Taheri, Yi Zhou, Dimitrios Tzionas, Yang Zhou, Duygu Ceylan, Soren Pirk, and Michael J. Black. GRIP: Generating interaction poses using latent consistency and spatial cues. In *International Conference on 3D Vision (3DV)*, 2024.

[41] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J. Black. DECO: Dense estimation of 3d human-scene contact in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[42] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. InterDiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[43] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. OakInk: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[44] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Grounding 3d object affordance from 2d interactions in images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[45] Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. G-HOP: Generative hand-object prior for interaction reconstruction and grasp synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[46] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[47] Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[48] Hui Zhang, Sammy Joe Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. ArtiGrasp: Physically plausible synthesis of bimanual dexterous grasping and articulation. In *International Conference on 3D Vision (3DV)*, 2024.

[49] Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[50] 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)*, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer:[Yes]

   Justification: The three contributions stated in the abstract and Sec. 1 — a hand-centric MANO vertex formulation, unified annotation across heterogeneous supervision, and a contact-gated VertexFormer — are supported by Sec. 3.1, Sec. 3.2, and the experimental results in Sec. 4 (Tables 1–5, Figs. 1 and 6).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Sec. 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Sec. 3 fully describes the unified annotation pipeline (tactile-to-vertex mapping, sensel-to-vertex projection, distance-based contact) and the VertexFormer architecture (vertex token encoding, spatial/temporal attention, contact-gated head). Sec. 4.1 reports datasets and splits, backbone choices (DINOv3 ViT-B/16, HaWoR), input resolution, clip length, transformer depth and width, optimizer, learning rate, batch size, training epochs, loss weights, and contact thresholds, which together are sufficient to reproduce the main results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release our implementation and the newly collected near-bare-hand evaluation data (23 clips, 6,310 frames), along with documentation and instructions for reproducing the reported experiments. Existing benchmark datasets are available from their original sources.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sec. 4.1 specifies all training and test details: dataset splits for OpenTouch, PVDB, DexYCB, ARCTIC, and MOW; backbone choices (DINOv3 ViT-B/16 and HaWoR, both frozen); input resolution and patch token count; clip length; transformer depth, hidden dimension, and number of heads; optimizer, learning rate, batch size, and training epochs; loss weights; and contact thresholds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: Results in Tables 1–5 are reported as single-run numbers (with a fixed seed) without error bars or confidence intervals, due to the computational cost of training the full model across multiple seeds on the combined dataset.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sec. 4.1 reports that training is conducted on a single RTX 3090 GPU with the specified optimizer, batch size, and number of epochs. Inference uses frozen DINOv3 ViT-B/16 and HaWoR backbones together with the lightweight VertexFormer head described in Sec. 3.2. It took about 8 hours for training.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The study uses publicly available research datasets and an additional near-bare-hand evaluation recorded by one of the authors. No external participants were recruited for the additional capture.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper does not include a dedicated broader impacts discussion. The proposed method targets hand-object pressure estimation as a perception task, with potential applications in AR/VR haptics, robot learning, and rehabilitation, and inherits the privacy considerations of general egocentric vision.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release new datasets or pretrained models that pose a high risk of misuse. The method builds on publicly available datasets and existing pretrained backbones.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets used in this work — datasets (OpenTouch, PressureVisionDB, DexYCB, ARCTIC, MOW) and pretrained models (DINOv3, HaWoR) — are properly cited in the references and used in accordance with their original licenses and intended research use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release the near-bare-hand evaluation data with documentation of the capture setup, data format, evaluation protocol, and usage conditions.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The near-bare-hand evaluation was recorded with one of the authors as the sole participant. No external participants were recruited, and no crowdsourcing, separate participant instructions, or compensation were involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The near-bare-hand evaluation was self-recorded by one of the authors, without external participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the core method.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.