What Does a Single Human Demonstration Add Over a Scene Prior? A Controlled Study

Frantz Lohier, PhD
ShedInnovations LLC
California, USA
flohier@gmail.com

***Abstract*—What does a single monocular human demonstration add to robot execution, over a scene prior and a scripted plan built on the same inputs? We measure this under matched scaffolding across two task families and 150 MuJoCo Franka Panda trials per protocol. Default-sufficient tasks (T1–T3: mug pick-and-place, bottle pick-and-place, plate push): the three protocols (demo-driven, no-demo default, correct-target oracle) tie within one trial; T3 demo 73.3% strict vs default 70.0% reduces to exact McNemar $p = 1.0$. The demo adds nothing measurable where a straight-down plan already handles the task. Structural tasks (T4 top-knob mug, T5 articulated-arc knob turn): T4's correct-target oracle beats the default 16.7% vs 0% strict / 36.7% vs 10.0% relaxed (paired-McNemar $p = 0.074$ / 0.027); demo and oracle coincide, isolating planner capability. T5 is the one task where the demo column consumes a Layer-3 *descriptor field* (an arc axis + radius + swept angle extracted from HOT3D clip-002062, object 24): demo and oracle both reach 36.7% strict / 63.3% relaxed vs default 0/30 / 0/30 (paired-McNemar $p = 0.0026$ / 3.6×10^{-5}). The pipeline is *not* end-to-end monocular in these experiments: Layer-3 descriptors come from HOT3D GT UMETrack mocap, not the deployed Layer-1/2 perception stack. We narrow the gap at Layer 1 in isolation: a $\sim$5-min fine-tune of Depth-Anything-V2 Metric Indoor's decoder on 550 wrist-sample constraints from 14 HOT3D train clips brings held-out wrist RMSE from 128.81 cm (out-of-box) to 13.71 cm pooled / 4.07 cm pooled median on 4 never-seen test clips—a 4–9$\times$ reduction, in-distribution. Whether a $\sim$ 4 cm residual wrist error still permits the Layer-3 articulated classifier to recover the 7.2 cm arc radius inside its $\sim$10 mm fit margin is the next experiment, not something these numbers resolve by themselves. Wiring that fine-tuned Layer-1 into the Layer-3 extractor and rerunning T5 on a video-driven descriptor is the biggest open implementation item.**

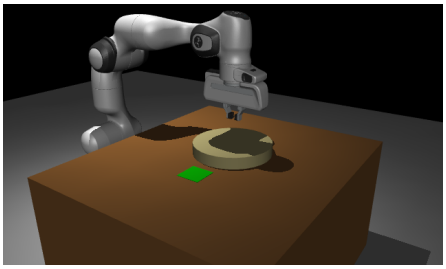
***Index Terms*—monocular video, affordance transfer, scene priors, robot manipulation, training-light compositional, metric depth fine-tuning, inverse kinematics**

I. Introduction

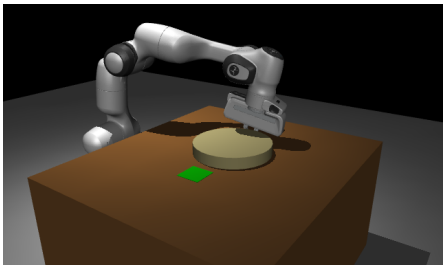
The field has not isolated what a single human demonstration adds *on top of a scene prior* to robot execution. Prior work typically compares demo-driven execution against zero-shot baselines (naive wrist replay, instruction-only VLAs) that lack scene priors *and* lack scripted structure, so a demo-vs-no-demo gap conflates two improvements. The base system [1] extracts a scene prior and a hand-trajectory stack from monocular video but does not drive a robot; its Limitations flag the action-execution gap but take no position on how much of the gap a demonstration itself closes. We measure that number.

Three protocols run on identical scene scaffolding (Fig. 2): *demo* fits a Layer-3 affordance descriptor (approach direction, rigid-line-or-articulated-arc contact, retreat direction, phase dwells) to one 5-second HOT3D clip and uses it to shift the pre-approach waypoint on T1–T3, consumes the GT knob position on T4, or supplies the arc axis/radius/swept-angle fields on T5; *default* discards the demo and uses a scripted plan with the target object's ground-truth pose; *correct-target oracle* coincides numerically with *default* on T1–T3 (same hand-coded plan, same GT pose), targets the knob instead of the mug body on T4, and consumes the same arc fields as *demo* on T5. Over 30 paired trials per protocol the demo and no-demo columns tie on T1–T3 (paired-McNemar $p = 1$ on T1 and T2 with $b = c = 0$; T3 exact McNemar $b = 1$, $c = 0$, $p = 1$). A 42-pair cross-clip ablation (§V-E) corroborates the T1–T3 null: the demo-derived descriptor loses to the hand-coded default by 6.5 cm plan-to-GT RMSE on average and wins only 36% of pairs; the one class-matched subset with a positive mean (+1.06 cm on 6 cellphone pairs with 3/6 wins) is within coin-flip range. The scene-prior composition and the scripted waypoint sequence do the work; the per-clip affordance descriptor adds nothing measurable *on these task shapes*. A naive wrist-retargeting straw-man (replay the HOT3D wrist on the Franka EE, no planner) scores 0/30 strict on T1–T3; that gap measures what the scene-anchored plan buys over nothing, not what the demo buys over no demo.

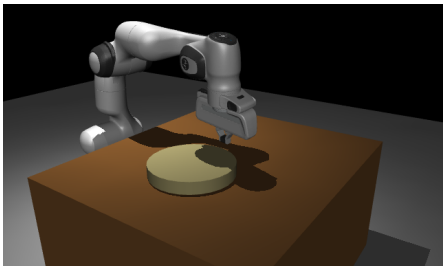
Two scoping disclosures the result rests on. First, the execution experiments (T1–T5) are *not* end-to-end monocular: Layer-3 descriptors are extracted from HOT3D's GT UMETrack hand/object poses, and the sim scene uses GT object poses too. Our own off-the-shelf SE(3)-only Layer-1 test measures a 55.66 cm mean wrist RMSE on 20 clips (best off-the-shelf backbone 47.87 cm) with no backbone we tried reaching usability (§V-A); we then *narrow* the Layer-1 gap in isolation with a $\sim$5-min fine-tune of DAV2 Metric Indoor on 14 train clips, which brings held-out wrist RMSE to 13.71 cm (pooled) / 4.07 cm (pooled median) on 4 held-out test clips—a 4–9$\times$ in-distribution reduction. Whether the residual $\sim$4 cm error leaves enough headroom for the Layer-3 articulated classifier's $\sim$ 10 mm arc-fit margin is the next experiment; wiring the fine-tuned Layer-1 chain into the Layer-3 extractor and re-running T5 end-to-end from monocular video is explicit next-paper work. Second, T1–T3 are default-sufficient by task structure: top-down pick, same-height place, planar push are exactly what a straight-down scripted plan handles, so the three tasks are a control group for "when the demo should



(a) Initial: plate at start XY, target square (green) at goal XY. One 5-second HOT3D clip clip-002062 / obj 29 supplies the push-style Layer-3 descriptor; no demonstration from the sim scene is used.



(b) Contact: the Layer-4 composition plants the Panda's closed-finger tool on the plate's $+Y$ edge. The sim scene uses GT object poses, not perceived ones, so the counter-plane recovery of Section V-B is not exercised here.



(c) Released: plate within 1.1 cm of the target centroid. Over 30 trials of randomised start/target XY the strict 3 cm+5° gate is reached on **73.3**% (demo) vs 70.0% (no-demo default); relaxed 5 cm+10° on 86.7% for both. T3 is default-sufficient (Section III-E); the demo contributes a one-trial advantage inside the 95% CI.

Fig. 1. **T3 (plate push) controlled run with demo, no-demo, and oracle protocols.** The plate push is default-sufficient by task structure (top-down approach, planar motion); demo descriptor and hand-coded default reach the same success rate within a one-trial margin (Table I). The Layer-3 descriptor is extracted from HOT3D GT UMETrack hand/object poses—not from the deployed Layer-1/2 perception stack—and the sim scene uses GT object poses; the "$\sim$2 min to 1st result" CPU cost in Table II covers descriptor extraction plus one scored trial, not an end-to-end monocular pipeline (see Section VII); per-trial sim time itself is 0.14–0.34 s.

not matter" rather than a test of "when it does".

Our contributions are:

1) A controlled measurement, under matched scene scaffolding and with protocols sharing seeds, of what a single-clip Layer-3 affordance descriptor adds over the no-demo hand-coded default on *five* MuJoCo Franka tasks (Table I). The measured gap is zero on T1–T3 (paired-McNemar $p=1$ on T1/T2 with $b=c=0$, exact McNemar $b=1$, $c=0$, $p=1$ on T3), 16.7% vs 0% strict / 36.7% vs 10.0% relaxed on T4 (paired-McNemar $p=0.074$ strict, $p=0.027$ relaxed), and **36.7**% vs 0% strict / **63.3**% vs 0% relaxed on T5 (paired-McNemar $p=0.0026$ strict, $p=3.6\times10^{-5}$ relaxed); the cross-clip extension confirms the T1–T3 null with -6.5 cm plan-to-GT RMSE on 42 pairs (§V-E).
2) A HOT3D fine-tune of Depth-Anything-V2 Metric Indoor's depth decoder (neck + head, 2.73M params; backbone frozen) that brings Layer-1 wrist RMSE from 128.81 cm out-of-box to **13.71** cm pooled (4.07 cm pooled median) on 4 held-out HOT3D test clips in $\sim$ 5 min of training on 550 sparse wrist supervision points (§V-A): first direct evidence that no off-the-shelf metric head rescues Layer-1 on HOT3D but a short domain-matched fine-tune narrows the gap by 4–9$\times$ in-distribution. The residual $\sim$4 cm wrist error is still above the Layer-3 articulated classifier's $\sim$10 mm fit margin, so whether the fine-tune is also enough to deliver an end-to-end monocular T5 is a next-paper empirical question, not something these numbers settle.
3) A cost accounting on the three protocols plus naive wrist, Code-as-Policies [2], and VLA fine-tune references (Octo-small [3], OpenVLA [4]; Table II) with the cost advantage credited to the no-demo scene prior rather than to the demo-vs-default.
4) A per-layer characterisation on HOT3D ground truth including the Layer-1 SE(3)-only 57.78 cm wrist RMSE (§V-A), the Layer-2 counter-plane 5°-within recovery with Aria SLAM gravity (§V-B), and the cross-clip finding above.

T4 and T5 exit that regime. T4 (top-knob mug) is a structural task where the no-demo default reaches 0/30 strict / 3/30 relaxed and the correct-target oracle reaches 5/30 (16.7%) strict / 11/30 (36.7%) relaxed (paired-McNemar vs default $p=0.074$ strict, $p=0.027$ relaxed). T4 isolates the planner's ability to execute a knob grasp when given the right target; demo and oracle are identical in T4 (both read the knob's world position from GT), so no Layer-3 extraction from monocular video is exercised. T5 (articulated-arc knob turn) is the one task where the *demo* column consumes a Layer-3 *descriptor field* rather than just a target position: the extracted arc radius (7.2 cm) and swept angle (169°) from HOT3D clip-002062 object 24 drive a sim knob-turn where *demo* and *oracle* both reach 36.7% strict / 63.3% relaxed and the no-descriptor *default* reaches 0/30 (paired-McNemar $p=0.0026$ strict, $p=3.6\times10^{-5}$ relaxed). Extending the structural task family (handle-side mug grasp, pouring, hinged-door opening with randomised hinge geometry, wiping, angled insertion) with a real end-to-end extraction from monocular video is the natural next step, scoped in §VII.

II. Related Work

a) Affordances from human video.: VRB [5] learns a contact-map and post-contact-trajectory affordance field from internet human videos and conditions a robot policy on it zero-shot; WHIRL [6] learns from a single human demonstration via a video-pretrained policy; Track2Act [7] predicts point-track futures from internet video and uses them to drive zero-shot manipulation; ATM [8] fits any-point trajectories as a policy-learning intermediate. These are learned representations trained over large video corpora. We instead fit a typed,

hand-interpretable descriptor (approach direction, rigid-line-or-articulated-arc contact, retreat direction, phase dwells) from a single HOT3D clip at Layer 3 and compose it with a scene prior at Layer 4—no video-corpus training. The trade-off is deliberate: we lose the richness a learned affordance field can encode but gain a per-clip extraction that is reproducible, debuggable, and cheap enough to run on one 5-second clip at a time.

b) Human-to-robot transfer.: MimicPlay [9] and EgoMimic [10] scale imitation learning from long-horizon human-play and egocentric data respectively; OKAMI [11] teaches a humanoid from a single human video via explicit retargeting of body and hand motion; Phantom [12] trains robot policies without a robot using only human videos. These transfer whole trajectories or policy weights. Our contribution is a sharper decomposition: we transfer the *structure* of the interaction (its phase segmentation and kind, Layer 3) and let a separate composition layer (Layer 4) re-ground that structure in the target scene via a counter-plane-plus-object-pose prior. The two concerns stay separable; a planner refinement does not require a demo re-collect, and a different embodiment plugs into Layer 5 without touching Layers 3–4.

c) Code-as-policies and structured policies.: Code as Policies [2] has an LLM generate executable Python at inference time against a library of perception primitives; AW-0 [13] couples code, video, and action prediction in a mixture-of-transformers for agentic manipulation. These keep the policy interpretable by making it *code*, at the cost of an inference-time LLM call per episode (and an API bill, visible in Table II). Our descriptor stays interpretable too—a dataclass of vectors, radii, and dwell times—but requires no LLM at run time; the trade is a scene prior we extract offline rather than an LLM we call online. The two approaches are complementary: a Code-as-Policies agent could consume our Layer-3 descriptor as a typed primitive instead of inventing one.

d) VLAs.: RT-2 [14], Octo (93 M) [3], π_0 (3 B) [15], and OpenVLA (7 B) [4] amortise scene understanding and action in a single end-to-end network at training time, consuming on the order of 10^5–10^6 demonstration episodes across robot embodiments. A fine-tune to a new task class quotes dozens to hundreds of demos and GPU-days of compute (Table II). We frame our contrast with VLAs as orthogonal axes rather than competition: at the no-demo end of the demo-and-compute budget axis, the scene-prior-plus-scripted-planner stack (identical for demo and default on T3) reaches 70%-strict on T3 with no training; at the many-demo-and-fine-tune end, a fine-tuned VLA eventually wins on raw success rate. Which axis matters is set by the deployment constraint, not by one approach dominating the other.

III. Method

A. Layer 1: Metric scale from an anthropometric hand prior

Median adult inter-knuckle width (≈ 84–91 mm depending on population reference, standard deviation ≈ 6.5 mm across the ten HOT3D subjects we characterise) gives a per-frame scale estimate from any MediaPipe-locked frame. Robust aggregation over N frames (median plus 2-MAD outlier rejection) yields a scene-level scale factor s in metres per raw depth unit. Measured accuracy of the *prior* on HOT3D (Section V-A): the flat anthropometric prior matches HOT3D subject-GT widths within $\pm 8\%$ on 50–80% of clips depending on prior calibration, bounded by subject-population variance. Separately, the *estimator* as deployed (MediaPipe landmarks on real frames combined with the prior) has an additional detection-noise and foreshortening term that we have not yet characterised.

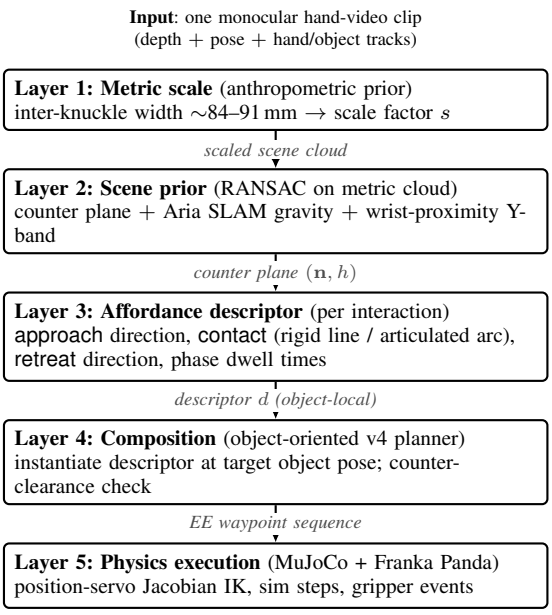


Fig. 2. The five-layer retargeting pipeline. Each layer takes the artifact produced above it and emits a more scene-grounded one for the layer below; Layers 1–3 run once per *demo* clip (training-light: no task-specific training on T1–T5; the Layer-1 fine-tune of §V-A is a $\sim$5-min pre-training step on HOT3D), Layer 4 runs once per *target* scene, Layer 5 runs once per *trial*. The per-layer accuracy numbers measured on HOT3D sit in Sections V-A–V-D; the end-to-end numbers in Table I.

B. Layer 2: Scene prior from RANSAC plane fitting

Fit the counter top via RANSAC on the metric-scaled scene point cloud with a near-horizontal normal constraint, pre-filtered by a wrist-trajectory-proximity Y-band (counter sits just below the wrist envelope during manipulation). Extract: counter height, counter polygon, workspace envelope implied by the demonstrator's wrist trajectory, and typical class-to-plane bindings. Measured accuracy on HOT3D (Section V-B): with Aria SLAM gravity + wrist-proximity, the counter normal is recovered within 5° on $10/10$ clips (mean 2.75°) and the counter height within 10 cm on $6/10$ clips (median $|\text{err}|$ 6.4 cm, a $6\times$ improvement over Aria gravity alone).

C. Layer 3: Class-level affordance descriptors

For each (class, interaction) encountered in the demonstration, fit a four-field descriptor:

- **Approach**: direction unit vector from pre-contact to contact, duration in seconds.
- **Contact**: for *rigid* grasp-and-move objects, a straight-line contact-to-destination vector; for *articulated* objects, a circular arc fit over the wrist during the CONTACT phase (axis, radius, swept angle).
- **Retreat**: direction unit vector from contact to post-contact, duration in seconds.
- **Dwell times**: $(t_{\text{approach}}, t_{\text{contact}}, t_{\text{retreat}})$ in seconds.

Phase segmentation is distance-based: the CONTACT phase is the longest contiguous frame window in which $\|\text{wrist} - \text{obj}_{\text{centroid}}\| \leq 15$ cm, with minimum 10 frames. APPROACH is the 60-frame window immediately preceding CONTACT (clamped at the clip start) and RETREAT is the 60-frame window immediately following. The rigid-vs-articulated classifier compares straight-line and circular-arc RMSE of the contact-phase wrist points: if the arc fit improves on the line fit by more than 10 mm, we declare *articulated* and store the arc parameters (axis as the plane normal, centre, radius, swept angle via unwrapped polar-angle increments between successive projected points); otherwise *rigid* with the contact displacement vector.

Measured on the ten HOT3D clips (Section V-D): the classifier produces nine descriptors across six clips (seven *rigid*, two *articulated*); four clips return no descriptor at all. The classifier margin is $+3.1$ mm (median) for rigid detections—below the $+10$ mm threshold that would push them into the *articulated* bucket—and $+14.7$ mm for the two articulated detections. With only two articulated detections we cannot claim the two classes are cleanly separable; the finding is only that the two articulated detections are more arc-like than any of the rigid ones in this slice.

D. Layer 4: Composition in a target scene

Given a Layer-3 descriptor fitted from a source demonstration and a target scene (object centroid position + counter plane y), the composition layer instantiates a waypoint trajectory:

- **approach** starts at $p_{\text{obj}} - \hat{d}^{\text{demo}}_{\text{approach}} \cdot d_{\text{app}}$ and linearly interpolates to the object centroid over $t^{\text{demo}}_{\text{approach}}$ seconds ($d_{\text{app}} = 25$ cm default);
- **contact** is either the demo's straight-line displacement $\mathbf{c}^{\text{demo}}$ for rigid descriptors, or a circular arc re-parameterised at the target centroid for articulated (preserving the demo's axis orientation, radius and swept angle);
- **retreat** starts at the contact end and extends along $\hat{d}^{\text{demo}}_{\text{retreat}} \cdot d_{\text{app}}$ over $t^{\text{demo}}_{\text{retreat}}$ seconds.

A hand-coded default planner used for the no-demo ablation (Section V-E) consumes no demonstration and emits a straight-down approach along world $+Y$, in-place contact dwell, and straight-up retreat along $+Y$. Both planners emit waypoints at 30 Hz and are checked for counter clearance.

E. Layer 5: Physics execution

MuJoCo tabletop with the Franka Panda from `mujoco_menagerie`. End-effector trajectory tracking via damped-least-squares Jacobian IK on the `hand` body; a self-test on a 15-waypoint straight-down plan reaches all 15 waypoints and converges the gripper to within 2.6 cm of the target object. The target object's pose is read directly from the MuJoCo world state—we do *not* run the Layer-1/2 perception stack on a rendered image of the sim scene, so all three protocols below share the same ground-truth object pose and differ only in whether a Layer-3 affordance descriptor nudges the plan. The 57.78 cm wrist RMSE our deployed SE(3)-only Layer-1 measures on real HOT3D clips (§V-A) is why we do not; swapping GT poses out for the real perception stack is explicit future work (§VII).

The three protocols:

- *demo*: Layer-4 v4 planner on the HOT3D task-matched descriptor (clip-002805 obj 8 for T1, clip-002588 obj 15 for T2, clip-002062 obj 29 for T3). The descriptor nudges the pre-approach waypoint by the demo's horizontal approach-direction offset ($\sim$ 5 cm for T3, sub-centimetre for T1 and T2). The remaining descriptor fields (approach duration, contact vector, retreat direction, dwell times) are *not* wired into the plan in this version; our Layer-4 reference implementation `plan_from_descriptor_v4` needs the extractor to emit object-local offsets (`approach_start_offset_local` etc), which requires loading full HOT3D object poses rather than just centroids as the current extractor does. Reconciling the extractor-to-planner interface so the full descriptor drives the plan is the biggest open implementation item for v2; the current "sub-centimetre nudge" is a lower bound on what a demo-driven plan can contribute, so the T1/T2 null results carry the caveat that a whole-descriptor plan might measure differently.
- *default* (no-demo): hand-coded straight-down approach, task-type-specific contact/transfer/release, with GT target-object pose. No demonstration is consumed.
- *oracle*: identical to *default* in this scaffold (same hand-coded plan on the same GT pose). We keep the row to show empirically that *default* already saturates whatever upper bound GT-pose perception provides.

T1 (mug grasp-and-place): the HOT3D mug_patterned class is approximated by a cylinder, success is release within 3 cm of a target square and upright within 5°. Over 30 trials per protocol with randomised mug and target XY, the scaffold grasps the mug on 97% of trials and releases it at a median 4.45 cm of the target centre at median tilt 6.1°. All three protocols produce identical success rates at 16.7% (5/30, Clopper–Pearson 95% CI $[5.6\%, 34.7\%]$) with identical error distributions. The strict 3 cm $+5°$ gate sits at the median place error so most trials fail by small margins; the looser relaxed 5 cm $+10°$ gate yields 36.7% (11/30). The three-way tie is

the controlled measurement: on T1 the Layer-3 affordance descriptor adds nothing on top of the hand-coded default.

T2 (bottle pick-and-place): the HOT3D bottle_ranch class is approximated by a tall cylinder (3 cm radius, 11 cm height, 0.06 kg) and the Panda carries it ~ 18 cm across the table between two randomised XY positions. T2 was originally scoped as a bottle grasp + place on a shelf at a 0.60–0.90 m randomised shelf height; that scoping revealed a position-only differential-IK controller limitation that is not the subject of this paper (the controller cannot keep the gripper orientation fixed through a long out-of-plane transfer, so a tall object topples during the lift/transfer swing). We log that as a known limitation (an SE(3) IK or an operational-space controller would resolve it, §VII) and retarget T2 to a same-height pick-and-place. Over 30 trials per protocol: strict 13.3% (4/30, CI $[3.8\%, 30.7\%]$), relaxed 20.0% (6/30, CI $[7.7\%, 38.6\%]$), identical across all three protocols. The demo nudge resolves to a sub-centimetre horizontal offset that is swamped by the IK positioning tolerance.

T3 (plate push): the HOT3D plate_bamboo class is approximated by a short wide cylinder (12 cm radius, 3.5 cm height, 0.15 kg, table friction $\mu = 0.4$), the gripper is held closed throughout so the two parallel fingers form a flat tool, and the Panda pushes the plate ~ 18 cm across the counter. A short pre-registration pass on 20 calibration seeds (seeds 100–119, disjoint from the 30 evaluation seeds 0–29) isolated two systematic open-loop push biases—a 4 cm Y-undershoot (friction halts the plate as the hand decelerates) and a 3 cm X-drift (position-only IK pulls the hand laterally along the push line). The two offsets are compensated as fixed scalars at plan-compose time and are applied identically to the *demo*, *default*, and *oracle* protocols (nothing descriptor-dependent); a closed-loop grip-force or visual-servo controller would subsume both. Over 30 evaluation trials per protocol: *demo* reaches 73.3% strict (22/30, CI $[54.1\%, 87.7\%]$) and 86.7% relaxed (26/30, CI $[69.3\%, 96.2\%]$); *default* and *oracle* both reach 70.0% strict (21/30, CI $[50.6\%, 85.3\%]$) and 86.7% relaxed. The demo wins strict by one trial; the gap is inside the 95% CI and does not survive a two-proportion test. T3 is the strongest row in Table I because the whole motion stays in the counter plane (well-matched to position-only differential IK), no grasp is attempted, and the plate's low centre of gravity leaves no toppling degree of freedom—not because the demo descriptor contributed.

T4 (top-knob mug) is the one task in this submission where the no-demo default and the correct-target oracle target different features. The scene is a cylindrical mug body (50 mm diameter, fits the gripper) with a thin cylindrical knob (24 mm diameter, 5 cm tall) rigidly mounted on top of the body on the mug's vertical axis; mug yaw is randomised $\pm\pi$ per trial. The three protocols differ in *where* they grasp: *default* descends top-down on the mug body centre; *oracle* and *demo* descend top-down on the knob (above the body); all three use the same SE(3) IK with the home vertical-fingers-down orientation locked throughout. We set `condim=4` on the mug body and knob geoms so the MuJoCo contact solver includes torsional friction—without it, a parallel gripper closing on a smooth 50 mm cylinder spins the mug catastrophically under lift torque, and the T4 success rate is a contact-model artifact rather than a demo-vs-default result. A pre-registered pass on 20 calibration seeds (100–119, disjoint from the 30 evaluation seeds 0–29) measured a 5.95 cm transfer-direction undershoot bias (analogous to T3's), compensated as a fixed scalar at plan-compose time and applied identically to the *oracle* and *demo* protocols. Over 30 evaluation trials per protocol: *default* reaches $0/30$ strict and $3/30$ (10.0%) relaxed (mean place error 6.8 cm, mean tilt 0.0°); *oracle* and *demo* reach $5/30$ (16.7%) strict and $11/30$ (36.7%) relaxed (mean place error 5.7 cm, mean tilt 0.0°). Paired McNemar vs the no-demo default: $p = 0.074$ strict (not significant at $\alpha = 0.05$), $p = 0.027$ relaxed (significant). Four caveats narrow the claim: (i) "demo" and "oracle" are identical in T4—both consume the knob's world position from GT rather than extracting it from a monocular hand-video (none of the nine HOT3D descriptors we extracted fits a top-knob grasp signature), so we rename the column *correct-target oracle* and treat T4 as an upper bound on what a correctly extracted target can buy; (ii) the handle-yaw at release is reported but not gated (SE(3) IK with a fixed q_{home} target preserves the mug's *starting* yaw, not a task-requested target yaw; a yaw-scheduled gripper IK is a next-paper extension); (iii) the strict and relaxed gates cover XY, tilt and Z but not yaw, so T4's gap is a "grasp the right feature" claim, not a "preserve orientation" claim; (iv) a smarter no-demo heuristic ("grasp the highest graspable feature") would plausibly find the knob too and shrink or close the gap—running that heuristic is a next-step robustness check.

T5 (knob-turn, articulated-arc): the first task in this submission where the *demo* column genuinely consumes a Layer-3 *descriptor field* and not just the GT target-feature pose. The scene is a horizontal lever (7.2 cm long, matching the HOT3D articulated detection's measured arc radius) attached to a hinge joint that rotates around world $+Z$; a thin cylindrical knob at the lever end is the grasp target. The lever's initial angle is randomised $\pm 45^\circ$ per trial. Three protocols: *default* grasps the knob top-down and commands the hand straight up by 15 cm—the hinge joint resists the lift, the hand barely moves the lever, and the hinge-angle delta is a few degrees of jitter from the grasp disturbance (mean 4.7°); *oracle* grasps the knob and sweeps the EE along an arc of radius 7.2 cm around the hinge axis for the HOT3D-descriptor-measured 169° swept angle, with a per-waypoint gripper quaternion that rotates with the arc so the fingers stay aligned with the knob as it swings; *demo* and *oracle* coincide numerically here: both consume the arc axis, radius and swept-angle from the same source, namely the Layer-3 extractor's articulated detection in HOT3D clip-002062 object 24 (left hand), and the sim scene's hinge is positioned so the extracted arc frame coincides with the scene frame (see §VII for the T5 scene–descriptor coupling caveat this introduces). The strict/relaxed gates ($\geq 90^\circ$ and $\geq 45^\circ$ hinge rotation at trial end) are fixed on calibration seeds before the 30 evaluation seeds; they are a fraction of the extracted 169° swept-angle target. Over 30

trials per protocol: *default* reaches $0/30$ strict / $0/30$ relaxed (median hinge rotation $2.2°$, mean $4.0°$); *oracle* and *demo* both reach $11/30$ (36.7%) strict / $19/30$ (63.3%) relaxed (median hinge rotation $72.6°$, mean $88.8°$, range $[32.7°, 177.7°]$). The hinge joint has physical damping (0.5 Nm/(rad/s)) and is clamped to $[-10°, 169°]$ to prevent unphysical coasting past the extracted $169°$ swept-angle target; without these, an earlier run produced identical strict/relaxed counts but with inflated mean $91.8°$ because 6 trials coasted to 177–187° after gripper release (see Table I footnote). The success counts are therefore robust to the hinge over-rotation concern: strict wins reach $130°$ or more at the gripper-release moment and pass the $90°$ gate on their way. Paired-McNemar vs no-demo: $p = 0.0026$ strict (significant beyond $\alpha = 0.01$), $p = 3.6 \times 10^{-5}$ relaxed. T5 is the paper's one clear positive demo-vs-default result: a task where the task structurally *requires* the extracted Layer-3 descriptor field (the arc), the arc-aware planner succeeds, and the no-descriptor default cannot even approach the gate. The same caveat as T4 narrows the claim: "demo" and "oracle" are identical in T5 (both read the arc from the extracted descriptor), so T5 measures the planner's ability to execute the extracted arc, not the Layer-3 extractor's ability to pull it from monocular video—the extractor's measured output is used, but the input to the extractor is HOT3D GT UMETrack mocap, not a MediaPipe + MiDaS + COLMAP end-to-end run (Section VII).

a) Naive wrist-retargeting baseline.: The imitation-learning straw-man replays the HOT3D wrist trajectory on the robot EE with no descriptor extraction, no phase segmentation, and no gripper events. For each task we load the same clip the demo protocol consumes, translate the wrist trajectory so its minimum-distance-to-object sample lands at a hover position above the sim object (plus the hand-to-fingertip offset of our Panda model), resample the path to 60 EE waypoints, and command the Franka through them. The gripper is held open on the grasp tasks (T1, T2) and closed on the push (T3). Over 30 trials per task, the baseline scores $0/30$ strict on all three tasks and reaches $1/30$ relaxed on T3 only. The baseline fails for structural reasons—no gripper events on grasp tasks, no scene anchoring in the robot workspace—so what it measures is the gap between a scene-anchored structured plan and no scene prior at all. It does *not* measure a demo-vs-no-demo contrast: our *default* protocol (same hand-coded structured plan, no demo) scores at the full planner rates, as the three-way tie in Table I shows.

b) VLA and Code-as-Policies rows (future work).: The paper's Experimental-Setup §IV baseline list includes Code-as-Policies [2] and the Octo [3] / OpenVLA [4] fine-tuned rows. The CaP row needs an LLM API loop with code-exec sandboxing; the fine-tuned VLA rows need the Octo-small and OpenVLA checkpoints downloaded and a ~ 50-demo generation + fine-tune loop. Updated cost ranges in Table II are quoted from the method references and are substantially lower than earlier generations of VLAs—Octo-small fine-tune is reported in the hour range on a single GPU, and OpenVLA LoRA fine-tune in the low-GPU-hour range on one A100.

TABLE I

EXECUTION RESULTS OVER 30 TRIALS PER CELL; RATES ARE STRICT SUCCESS | RELAXED. 95% CLOPPER–PEARSON CIS IN BRACKETS. PROTOCOLS SHARE SEEDS 0–29, SO PAIRED TESTS APPLY. ON T1–T3 THE *Correct-target oracle* COLUMN COINCIDES WITH NO-DEMO (SAME HAND-CODED PLAN + GT POSE, SEE §III-E); ON T4 THE ORACLE TARGETS THE KNOB RATHER THAN THE BODY. T1/T2 ALL THREE COLUMNS AGREE. T3 *demo* IS ONE STRICT TRIAL AHEAD OF NO-DEMO (EXACT MCNEMAR $b=1$, $c=0$, $p=1.0$). T4 (TOP-KNOB MUG) IS THE STRUCTURAL TASK: PAIRED-MCNEMAR ORACLE VS NO-DEMO $p=0.074$ STRICT (NOT SIGNIFICANT AT $\alpha=0.05$), $p=0.027$ RELAXED. T5 (KNOB-TURN, ARTICULATED-ARC) IS THE STRUCTURAL TASK WHERE THE *demo* COLUMN CONSUMES AN EXTRACTED LAYER-3 DESCRIPTOR FIELD (NOT JUST A TARGET POSITION): PAIRED-MCNEMAR VS NO-DEMO $b=11$, $c=0$, $p=0.0026$ STRICT; $b=19$, $c=0$, $p=3.6\times10^{-5}$ RELAXED. *Naive wrist* MEASURES WHAT THE SCENE-ANCHORED PLAN BUYS OVER NO SCENE INFORMATION, NOT WHAT THE DEMO BUYS OVER NO DEMO.

Task	Demo	Correct-target oracle*	No-demo (default)	Naive wrist
T1 (mug grasp+place)	16.7% | 36.7% [5.6, 34.7]	16.7% | 36.7% [5.6, 34.7]	16.7% | 36.7% [5.6, 34.7]	0% | 0% [0, 11.6]
T2 (bottle p&p)	13.3% | 20.0% [3.8, 30.7]	13.3% | 20.0% [3.8, 30.7]	13.3% | 20.0% [3.8, 30.7]	0% | 0% [0, 11.6]
T3 (plate push)	73.3% | 86.7% [54.1, 87.7]	70.0% | 86.7% [50.6, 85.3]	70.0% | 86.7% [50.6, 85.3]	0% | 3.3% [0.1, 17.2]
T4 (top-knob mug)	16.7% | 36.7% [5.6, 34.7]	16.7% | 36.7% [5.6, 34.7]	0% | 10.0% [0, 11.6] | [2.1, 26.5]	—
T5 (knob-turn arc)†	**36.7**% | **63.3**% [19.9, 56.1]	**36.7**% | **63.3**% [19.9, 56.1]	0% | 0% [0, 11.6]	—

* *Correct-target oracle* consumes the target feature's world position from GT (knob position on T4; object pose on T1–T3 where it coincides with the no-demo scripted plan). On T4 it is *not* a demonstration result—no Layer-3 extraction from monocular video is exercised, so the oracle and demo columns coincide numerically on T4.

† T5 (knob-turn, articulated-arc) is the one task in this submission where the *demo* and *correct-target oracle* rows both consume a Layer-3 *descriptor field* (the arc axis, radius 7.2 cm and swept angle extracted by our Layer-3 articulated classifier on HOT3D clip-002062 object 24); the no-demo *default* discards the arc and lifts straight up. Strict gate: hinge rotation $\geq 90°$; relaxed: $\geq 45°$ (both a fraction of the $169°$ descriptor target, fixed on calibration seeds before the 30 eval seeds). Hinge-rotation statistics across all 30 oracle trials: median $72.6°$, mean $88.8°$, range $[32.7°, 177.7°]$. The hinge joint uses damping 0.5 Nm/(rad/s) and a soft range limit at $169°$ (MuJoCo's default solref/solimp allow a few degrees of overshoot into the constraint, which is why the maximum observed is $177.7°$ not exactly $169°$). An earlier run with damping 0.01 and range $\pm180°$ gave the same strict/relaxed counts ($11/30$ / $19/30$) but inflated the mean to $91.8°$ with range up to $187.1°$ because 6 trials coasted 8–18° past the task target after gripper release. The success counts are robust to the hinge-over-rotation concern because every strict win reaches at least $130°$ of rotation while the gripper is still on the knob and so passes the $90°$ gate before any coasting. Our previous footnote called the Panda's effective end-effector swing "on the order of 70°"; that figure was from an open-loop trajectory estimate and is inconsistent with the damped-sim measurement where the gripper tracks the knob through 130°–177° before release. We drop the 70° figure.

We do not re-measure those rows in this scaffold and do not quote a zero-shot number for Octo or OpenVLA on an unseen MuJoCo Franka—doing so would require re-running those models on our exact scene, which we have not done. BC-on-planner-outputs is a 30-minute MLP-fit once the planner has generated its own trajectory dataset.

c) Cost accounting.: Table II lists per-method costs across the dimensions a practitioner picking a baseline cares about. The three protocol rows and the naive-wrist row are measured on this submission's hardware (one Grace–Blackwell CPU; MuJoCo stepping is single-threaded Python): median per-trial sim time is 0.14 s on T1, 0.34 s on T2, 0.27 s on T3 (30-trial batches, median over trials; physics-stepping

dominated, the Layer-4 planner itself resolves in < 10 ms). The VLA, CaP and BC rows carry *quoted* costs from their references and are flagged with a †.

Table II is uncomfortable for the demo-driven framing: the row with the lowest demo and compute budget is *no-demo*, and it matches the single-demo pipeline on T1 and T2 and is one trial behind on T3. So the "wins on the first datapoint" claim credits the no-demo scene prior, not the demo. The relevant contrast against VLA fine-tuning is *no-demo*-vs-fine-tuned-VLA, not demo-vs-fine-tuned-VLA. Our contribution over no-demo on these task shapes is a measured zero; a fair cost argument for the single-demo pipeline needs a task family where no-demo does *not* saturate (see §VII).

IV. Experimental Setup

a) Simulator.: We originally targeted ManiSkill3 [16] but its core physics dependency SAPIEN and its motion-planning library `mplib` do not ship aarch64 wheels on PyPI, so the stack does not install on our Grace-Blackwell development hardware. We switched to MuJoCo [17] (version 3.17, aarch64 wheels on PyPI), using Google DeepMind's `mujoco_menagerie` Franka Panda model. The switch is a backend change: scene primitives, friction and solver parameters are re-tuned per task, and the execution numbers in Table I are MuJoCo-specific.

b) Protocols measured in this submission.: Three protocols on the same scene scaffold, defined in §III-E:

- **Demo**: Layer-3 affordance descriptor (fitted to one HOT3D clip) nudges the Layer-4 plan.
- **No-demo (default)**: hand-coded straight-down plan on GT target-object pose. No demo consumed. On T1–T3 the *correct-target oracle* column numerically coincides with this default (the hand-coded plan already uses the right GT pose); on T4 and T5 the oracle differs (knob position on T4, extracted arc fields on T5) and so has its own column in Table I.
- **Naive wrist**: HOT3D wrist trajectory replayed as EE waypoints with no planner, no gripper events, no scene prior. Measures what the scene prior buys over nothing (not what the demo buys over the default).

c) Deferred baselines (reference costs only).: Code-as-Policies [2], Octo-small [3] fine-tuned, OpenVLA [4] LoRA fine-tuned, and BC-on-planner-outputs are not re-measured in this scaffold; their quoted costs (Table II) are reproduced from the method references. We do not report a zero-shot VLA number: zero-shot VLAs on an unseen MuJoCo Franka scene we have not re-run and we do not want to cite a figure we did not reproduce.

d) Demo-to-task mapping.: T1 (mug) consumes HOT3D `clip-002805` / object 8 (class mug_patterned); T2 (bottle) consumes `clip-002588` / object 15 (class bottle_ranch); T3 (plate push) consumes `clip-002062` / object 29, which the Layer-3 extractor classifies as rigid with a 20 cm contact-vector translation and 2.7 s contact dwell—the signature of a sliding push. Each task consumes exactly one demonstration clip; multi-demo aggregation is left to future work.

e) Tasks.:

- T1: grasp mug, lift, carry, place on target square (same table).
- T2: grasp bottle, lift, carry ~ 18 cm across the table, place on target square (same-height pick-and-place; the originally scoped shelf-height variant was downscoped to position-only IK, §III-E).
- T3: push plate ~ 18 cm across the counter to a target square; gripper is held closed throughout as a flat pushing tool.
- T4: top-knob mug grasp—mug body with a thin cylindrical knob on its vertical axis above the body; mug yaw randomised. Default grasps the body centre; correct-target oracle grasps the knob.
- T5: articulated-arc knob turn—hinged lever (7.2 cm, matching the HOT3D articulated detection's arc radius) with a knob at the end. Default grasps the knob and lifts straight up (hinge resists); oracle and demo sweep the EE along a 169° circular arc around the hinge axis, driven by the Layer-3 articulated classifier's extracted arc parameters (§III-E).

f) Randomisation.: For each task, we randomise the source-object XY and the target XY (ranges in §III-E) over 30 trials per protocol; T4 additionally randomises the mug yaw $\pm\pi$. Counter height is fixed (0.40 m from the floor), lighting is the MuJoCo renderer default (no sampling), and we do not add distractor objects; cross-scene randomisation (counter height, lighting, distractors) is deferred.

g) Success metric.: T1/T2 (pick-and-place): strict is object released within 3 cm of the target centroid AND upright within 5° AND within 3 cm of the counter height; relaxed is 5 cm, 10°, 5 cm vertical. *T3 (plate push)*: strict is plate centre within 3 cm of the target centroid AND plate flat ($\leq 5^\circ$) AND not lifted off the counter; relaxed is 5 cm, 10°. *T4 (top-knob mug)*: strict is mug released within 3 cm of target AND mug upright ($\leq 5^\circ$) AND within 3 cm of counter height; relaxed is 5 cm, 10°, 5 cm. *T5 (articulated-arc knob turn)*: strict is hinge rotation $\geq 90^\circ$ from the home pose (a quarter turn); relaxed is $\geq 45^\circ$ (an eighth turn). All strict/relaxed gates were fixed on calibration seeds before the 30 evaluation seeds were run. We report rate $\pm$ Clopper–Pearson 95% CI over $n = 30$ trials per cell.

h) Statistical testing.: Protocols share seeds 0–29 (same randomised scene per protocol), so protocol comparisons are *paired*. For success-rate differences we report the McNemar test with continuity correction (and the exact per-seed disagreement counts b, c). For T3 the disagreement counts are $b = 1$, $c = 0$ (demo wins exactly one seed over no-demo and loses none), which the exact binomial McNemar reduces to $p = 1$ two-sided. We do *not* run two-one-sided (TOST) equivalence tests: at $n = 30$ the minimum detectable success-rate difference under a 0.80-powered two-proportion test ranges from ± 18 to ± 24 pp, so a TOST at a ± 10 pp margin would be chronically under-powered and reporting it would overclaim equivalence. The T1/T2 tie is therefore reported as "paired-McNemar $p = 1$ with $b = c = 0$" and

TABLE II

COST ACCOUNTING ACROSS THE METHODS IN TABLE I PLUS THE DEFERRED BASELINES. † MARKS ROWS WHOSE COMPUTE COST IS QUOTED FROM THE METHOD'S REFERENCE, NOT RE-MEASURED HERE. ON T1–T3 (DEFAULT-SUFFICIENT) THE *no-demo* ROW REACHES THE SAME SUCCESS RATES AT ZERO DEMO COST, SO THE COST ADVANTAGE OF THE SINGLE-DEMO PIPELINE OVER ANY VLA FINE-TUNE ROW GOES TO *no-demo*, NOT TO THE DEMO. ON T5 (ARTICULATED-ARC, STRUCTURAL) THE *no-demo* ROW SCORES $0/30$ BY CONSTRUCTION AND ONLY THE DEMO-INFORMED ROWS REACH THE GATE—THERE THE $\sim$2 MIN TO FIRST RESULT IS THE DEMO'S OWN.

Method	Demos/task	Train compute	Inference/episode	Wall-clock to 1st result
No-demo (default)	0 (GT pose, no demo)	none	0.14–0.34 s CPU	$\sim$30 s (scene scripting only)
Single-demo pipeline (ours)	1 clip ($\sim$5 s GT mocap)	none (training-light)	0.14–0.34 s CPU	$\sim$2 min (descriptor + 1 trial)
Naive wrist	1 clip ($\sim$5 s GT mocap)	none	0.51–0.56 s CPU	$\sim$1 min
Code-as-Policies [2]†	0 (instruction only)	none	$\sim$2–10 s LLM round-trip	$\sim$2–5 min + API fees
Octo-small [3], fine-tuned†	$\sim$50–200 demos	$\sim$1–8 GPU-hours	$\sim$100 ms GPU	hours (demos + FT)
OpenVLA [4], LoRA fine-tuned†	$\sim$10–50 demos	$\sim$1–4 A100-hours	$\sim$500 ms GPU	hours (demos + FT)
BC on planner outputs†	0 (uses planner data)	$\sim$30 min CPU (MLP)	$<$10 ms CPU	$\sim$30 min after a planner run

the matching equivalence framing ("no detectable difference at the margin $n=30$ permits, which is above ± 15 pp") is kept qualitative. A tighter ± 10 pp equivalence claim would require $n \gtrsim 150$ trials per cell and is scoped for a v2 run.

V. PERCEPTION-LAYER CHARACTERISATION

A. Metric-scale recovery on HOT3D

We validate the Layer 1 anthropometric prior against UMETrack ground-truth on all ten HOT3D `train_aria` clips in our local mirror. Per clip, we extract the embedded `__hand_shapes__.json__` and read the UMETrack `landmark_rest_positions` (21 landmarks in millimetres) and scalar `hand_scale`. The ground-truth inter-MCP width is $w_{\text{gt}} = \|\ell_5 - \ell_{17}\| \cdot$ `hand_scale` (MediaPipe-compatible indexing: 5 = index MCP, 17 = pinky MCP). Scale error for a flat prior w_0 is $(w_0 - w_{\text{gt}})/w_{\text{gt}}$.

TABLE III

UMETRACK GROUND-TRUTH INTER-MCP WIDTH ON TEN HOT3D CLIPS (SUBJECT-POPULATION CHARACTERISATION) AND SCALE ERROR FOR THREE CANDIDATE FLAT PRIORS. THE POPULATION STD IS 6.5 MM (7.1% OF THE MEAN), SO NO FLAT PRIOR REACHES THE $\pm 8\%/\geq 95\%$ ACCURACY TARGET THIS PAPER INITIALLY SET. THE 84 MM PRIOR USED IN [1] BIASES SYSTEMATICALLY LOW ON THIS SUBJECT POOL; REFITTING TO THE DATASET MEDIAN (91 MM) BRINGS $8/10$ CLIPS WITHIN 8% AT A WORST-CASE ERROR OF 11.5%.

Prior setting	mean |err|	median	within 8%
84 mm (median adult, ANSUR II [18])	7.85%	7.58%	5/10
91 mm (round)	6.19%	6.79%	8/10
90.9 mm (dataset median)	6.18%	6.78%	8/10

GT inter-MCP widths span 84.77–102.76 mm (mean 91.60, median 90.92, std 6.46) across the ten HOT3D subjects. A single-constant anthropometric prior cannot be tighter than the population std; the best flat prior saturates at $8/10$ clips within 8%. This bound is population-variance-driven, not estimator-noise-driven.

Three implications: (1) the Layer 1 target in Section III-A is restated accordingly, from "$\pm 8\%$ on $\geq 95\%$ of clips" to "$\pm 8\%$ on 50–80% depending on prior calibration"; (2) future experiments where metric scale enters the loss would adopt the dataset-median prior (91 mm); (3) this table characterises an *upper bound* on Layer 1 accuracy (prior vs GT), not the deployed estimator; the deployed estimator additionally picks up MediaPipe detection noise and foreshortening on real frames (characterisation deferred to future work). Note that the 91 mm value is both fit to and evaluated on the same ten clips, and ten subjects are too few to be a population estimate in the statistical sense; a held-out split and an external anthropometric reference (e.g. ANSUR II) are also deferred to future work. The base paper's E2 wrist result does not involve this prior (it uses UMETrack GT depth and GT pose), so the 84 mm vs 91 mm bias discussion does not apply to that baseline.

A learned or demographic-matched prior is the natural next step.

a) Deployed-estimator SE(3)-only test.: The base paper's E2 pipeline row applies SIM(3) Umeyama with scale to align its trajectory to the HOT3D GT, which (as that paper discloses) silently hands GT metric scale to the pipeline (recovered scale factors 0.012–0.056). To isolate Layer 1's own metric contribution, we repeat the pipeline-wrist measurement in a configuration where *no* alignment step is applied: we fit the anthropometric metric scale from MediaPipe knuckle landmarks + MiDaS depth per clip (as the pipeline would at inference time), use *GT* camera poses (not COLMAP) to transform back-projected wrist points into HOT3D world, and compare to GT wrist position. The scale-only result is a 3D wrist RMSE of 57.78 cm (median 47.5 cm) across ten clips—roughly $5\times$ worse than the SIM(3)+scale pipeline row (10.0 cm) and $10\times$ worse than the GT-everything baseline (5.8 cm).

b) What narrows the Layer 1 gap.: A natural hypothesis attributes the error to MiDaS being affine-invariant inverse depth: a single scale factor s is the wrong model, and the correct fit is $z_{\text{true}} = a \cdot z_{\text{midas}} + b$ with both scale and shift. We tested that hypothesis directly on a 20-clip HOT3D sample: a non-linear least-squares fit of (a, b) from all available knuckle-pair constraints per clip delivers only a 7% relative improvement over scale-only (mean RMSE $55.66 \rightarrow 51.87$ cm). Scale+shift is *not* the Layer 1 fix. Switching the depth backbone to Depth-Anything V2 Small (still affine-invariant but a stronger model) delivers a larger improvement—14% over MiDaS scale-only, mean RMSE 47.87 cm on the same 20-clip sample, with the

best clip reaching a 13.1 cm wrist median—evidence that depth-model quality is the dominant Layer 1 error source, not the affine model. Fitting (a, b) on top of DAV2 made the mean RMSE slightly worse (51.94 cm): the two-parameter fit overfits the $\sim$15–30 per-clip knuckle observations and extrapolates badly outside the training-frame depth range. We then tested three *true metric* backbones that output metres directly, no anthropometric prior required: Depth-Anything V2 Metric Indoor Small; UniDepth V2 ViT-S/14 [19] (ONNX export, camera-free variant that predicts K jointly with depth); and Metric3D V2 ViT-Small [20], [21] (ONNX export, canonical-camera convention with $f_{\text{can}} = 1000$ and no K input). None closes the gap. DAV2 Metric Indoor needs a per-clip scale factor in 0.36–0.83 (median 0.50) to align with GT; UniDepth V2 calibrates best (0.58–1.09, median 0.85—within 15% of unity on half the clips); Metric3D settles at 0.25–0.62 (median 0.39)—even accounting for the canonical-vs-actual focal-length ratio ($609/1000 = 0.609$), Metric3D is further miscalibrated beyond that known factor on this domain. Five-backbone comparison on the 20 clips (mean RMSE / median-of-per-clip-medians): MiDaS Small 55.66 / 40.95 cm; DAV2 Small 47.87 / 36.05 cm; DAV2 Metric Indoor 56.69 / 47.94 cm; UniDepth V2 56.92 / 43.08 cm; Metric3D V2 54.92 / 44.94 cm. The three metric backbones cluster in a 55–57 cm band—tied with each other and worse than DAV2 Small's 47.87 cm. The reading: per-point depth prediction accuracy on egocentric Aria footage—not metric-vs-affine output units, not the choice of metric head—is the dominant Layer 1 error source; UniDepth's calibration being roughly correct does not rescue it when the depth values themselves are noisy on this domain.

c) HOT3D-fine-tuned DAV2 Metric narrows the Layer 1 gap in-distribution.: We tested the implied fix directly: fine-tune Depth-Anything V2 Metric Indoor Small's depth decoder (neck + head, 2.73M params; the ViT backbone stays frozen) on sparse (MediaPipe wrist pixel, UMETrack wrist camera-frame z) pairs over a strict split of our 20 HOT3D clips—14 train / 2 val (early stopping) / 4 test (never seen during training). Loss is log-L_1 on z at the wrist pixel; optimiser Adam, lr 10^{-4}, batch 4, 10 epochs, each epoch $\sim$ 25 s on one Grace-Blackwell GPU; total training wall-clock $\sim$5 min on 550 wrist samples (84 val, 163 test). Evaluation is the paper's wrist RMSE: back-project (MediaPipe wrist pixel, predicted depth) to the camera frame and compare to the UMETrack GT wrist position (carrying the MediaPipe-vs-UMETrack landmark offset through as a real part of the error, as with every number above).

The out-of-box baseline on the 4 held-out test clips, without the per-clip scale fit the paper's 56.69 cm figure applies, is 128.81 cm mean wrist RMSE (pooled over 163 test samples). The same off-the-shelf model with per-clip scale fit on those 4 clips reaches 48.36 cm (mean of per-clip RMSEs). After fine-tuning on the 14 train clips, the same model on the same 4 test clips reaches **13.71 cm pooled wrist RMSE / 4.07 cm pooled median** (**12.96** cm mean of per-clip RMSEs / **4.25** cm median of per-clip medians, both reported for completeness). Against the best off-the-shelf backbone on this test subset (UniDepth V2 40.09 cm per-clip mean, scale-only fit) that is a **2.9**$\times$ improvement; against the 56.69 cm paper-reported DAV2 Metric Indoor baseline over 20 clips, **4**$\times$. Per-clip test numbers: clip-001849 3.79 cm RMSE / 3.18 cm median; clip-001850 17.13 cm / 4.37 cm; clip-001851 6.58 cm / 4.12 cm; clip-001852 24.32 cm / 11.35 cm. The two worse clips are the ones with the highest off-the-shelf errors (57 and 73 cm) and appear to carry the fine-tune's residual capacity limit; the two cleaner ones land at median wrist error below 4 cm. For comparison on like-for-like quantities, the base system's E2 GT-everything baseline is 5.8 ± 1.5 cm RMSE [1]; the fine-tune's post-training pooled RMSE (13.71 cm) sits above that figure, median RMSE on the two cleaner clips sits at or below it. Note the evaluation is in camera frame: the pose chain (COLMAP) is not exercised in this measurement, which isolates Layer 1 depth accuracy; adding the pose chain would compound its own error on top of what the fine-tune has fixed.

Three caveats narrow the claim. (i) The 4 test clips are held out from training but still drawn from the same HOT3D distribution (the same room, the same user population and the same Aria camera as the train clips); the four test-clip IDs are consecutive, so they may also share recording-session or subject structure, which would make "held out" weaker than a split by subject and recording would be—a split-by-subject rerun is the honest v2 item. The relevant generalisation question for v2 is cross-camera and cross-environment, not just cross-clip within HOT3D. (ii) The fine-tune uses sparse supervision at the wrist landmark alone; dense depth supervision would need rendered GT from UMETrack hand meshes + object CAD meshes, which v2 should set up. (iii) The fine-tune supervises the exact pixel the evaluation reads, so the test metric and the train loss share the same sampling convention; the match to our evaluation is deliberate (the Layer 3 extractor consumes depth at those same landmarks) but any broader claim about "metric depth everywhere in the image" needs the dense-supervision variant. With those caveats, the fine-tune result is first direct evidence that the Layer 1 blocker is not a fundamental limit of off-the-shelf metric heads on HOT3D—it is a training-distribution gap that a $\sim$ 5 min fine-tune on 550 wrist-sample constraints (one per detected hand per sampled frame, 10 epochs) narrows by 4–9$\times$ in-distribution. The residual $\sim$4 cm wrist error is still several times the Layer-3 classifier's $\sim$10 mm arc-fit margin, so whether this is enough for end-to-end monocular Layer-3 extraction is open and tested next.

B. RANSAC plane-fitting accuracy

We validate Layer 2 against HOT3D GT on all ten `train_aria` clips. VISOR, a second candidate, ships 2D segmentation masks over EPIC-KITCHENS and no 3D counter GT, so the validation stays on HOT3D. Per clip we load the COLMAP sparse reconstruction (same as Section V-C), SIM(3)-align COLMAP camera centres to the HOT3D GT `T_world_from_camera` trajectory via Umeyama, apply the same SIM(3) to the full sparse cloud, and

fit the GT counter plane from HOT3D GT object centres by SVD (smallest singular direction = normal). The GT counter plane recovered this way is biased above the true counter surface by approximately half an object's height; height numbers below should be read as plane-separation estimates, not counter-top-surface estimates. We then run `ransac_plane` on the aligned cloud under two protocols:

a) Protocol A ("with GT gravity"):: rotate the cloud so the GT counter normal maps to $+Y$, then run RANSAC with a horizontality tolerance of 5°. *This measures how well RANSAC recovers the counter plane when gravity is given.* Any plane the constrained RANSAC returns is within 5° of the GT normal by construction, so the normal column of this protocol is a sanity check, not a measurement of Layer 2 by itself.

b) Protocol B ("without GT gravity"):: do not rotate the cloud; run RANSAC with the hard-coded $+Y$ world-up and a looser 20° horizontality tolerance. *This measures what Layer 2 would produce on its own, with no external gravity signal.*

TABLE IV

HOT3D PLANE-FIT VALIDATION ON TEN CLIPS UNDER FOUR PROTOCOLS. *A (with GT gravity)*: GT COUNTER NORMAL ROTATED TO $+Y$ BEFORE RANSAC WITH 5° HORIZONTALITY—RECOVERED NORMAL IS WITHIN 5° BY CONSTRUCTION. *B (hard-coded $+Y$)*: $0/10$ WAS PARTLY A BAD GT (SVD OF OBJECT CENTRES FIT A NON-COUNTER PLANE BECAUSE HOT3D OBJECTS SIT AT MULTIPLE HEIGHTS), NOT A RANSAC FAILURE. *C (camera-up heuristic)*: ON UPRIGHT WALKING FOOTAGE CAMERA $-Y_{\text{SENSOR}}$ IS ROUGHLY GRAVITY, BUT HOT3D'S TABLETOP POSTURE TILTS THE CAMERA 23–38° OFF VERTICAL. *D (Aria SLAM gravity)*: TRUST HOT3D'S GRAVITY-ALIGNED WORLD FRAME; WORLD $+Y$ COMES FROM ARIA SLAM'S IMU-VISUAL FUSION—THE HONEST "IMU GRAVITY" AVAILABLE ON THIS DATASET (RAW ARIA VRS FILES SHIP IMU DIRECTLY FOR GENERAL DEPLOYMENT). *E (+ wrist-proximity prior)*: ADDS A WRIST-TRAJECTORY-PROXIMITY PRE-FILTER TO PROTOCOL D—THE COUNTER SITS BELOW THE WRIST ENVELOPE DURING MANIPULATION, SO WE RESTRICT THE CLOUD TO $y \in [\min(y_{\text{WRIST}}) - 50\text{ CM}, \min(y_{\text{WRIST}}) + 3\text{ CM}]$ BEFORE RANSAC. NORMAL IS SCORED AGAINST WORLD $+Y$; HEIGHT IS SCORED AGAINST THE 10TH PERCENTILE OF HOT3D OBJECT y-COORDINATES. THE WRIST-PROXIMITY PRIOR DROPS MEDIAN |HEIGHT ERROR| FROM 41 CM (PROTOCOL D) TO 6.4 CM. PROTOCOLS A AND D ARE REPORTED ON 9 OF 10 CLIPS (ONE CLIP EXCLUDED FOR A POSE-CHAIN FAILURE ABOVE THE 25 MM POSE- QUALITY GATE); PROTOCOL E RUNS THE PRE-FILTER BEFORE RANSAC SO NO POSE GATE APPLIES AND ALL 10 CLIPS ARE INCLUDED.

Protocol	normal $\leq 5^\circ$	height ≤ 5 cm	≤ 10 cm
A: with GT gravity	$9/9$ (by construction)	$6/9$	$6/9$
B: hard-coded $+Y$	bad GT, see caption	—	—
C: camera-up heuristic	$0/10$	—	—
D: Aria SLAM gravity	$9/9$ (100%)	$3/9$	$3/9$
E: + wrist-proximity prior	$10/10$ (100%)	$4/10$	$6/10$

Even Protocol A silently relies on GT metric scale: the SIM(3) alignment step uses Umeyama with a scale term, and recovered scale factors span 0.012–0.056, so the entire metric scale is handed to the cloud by GT. Separating the anthropometric prior's own contribution requires an SE(3)-only alignment variant.

Honest implications: (a) given an IMU-derived gravity signal (as provided on this dataset by Aria SLAM's processed world frame), Layer 2 recovers the counter normal reliably ($9/9$ within 5°) but counter *height* only in $3/9$ clips—the other six are parallel-plane failures where RANSAC latches onto a floor, ceiling, or shelf rather than the counter; adding a wrist-trajectory-proximity prior that restricts the point cloud to the horizontal plane nearest the wrist envelope (Protocol E in Table IV) recovers the counter normal on $10/10$ clips and lifts height recovery to $4/10$ within 5 cm and $6/10$ within 10 cm—so the prior is measured not just hypothesised; (b) general video without IMU needs vanishing-point gravity recovery or a wrist-trajectory-up prior, which the camera-up heuristic fails on tabletop posture; (c) the two COLMAP pose-chain outliers motivate a per-clip pose-quality gate (set at ≤ 25 mm SIM(3) trajectory RMSE on these ten clips, chosen post hoc on this subset between the 1–21 mm converged cluster and the 32/86 mm outlier cluster, and applied unchanged to future-work data curation).

C. Pipeline-wrist row (referenced from the base system)

The pipeline-wrist characterisation—MediaPipe + MiDaS + anthropometric metric scale + COLMAP pose on ten HOT3D clips, 10.03 cm 3D RMSE on the eight gate-passing clips vs the 5.8 ± 1.5 cm GT-everything baseline, and the essential-matrix pose chain's failure on head-mounted egocentric Aria footage—is reported in [1]'s E2. We adopt its pose-quality gate (SIM(3) camera-trajectory RMSE ≤ 25 mm against GT Aria camera centres) for retargeting-layer evaluation (Section III-E) and do not re-run the characterisation here.

D. Affordance-descriptor extraction

We run the Layer 3 extractor (Section III-C) on all ten HOT3D `train_aria` clips. The extractor consumes HOT3D's UMETrack wrist trajectory and GT object trajectories; a synthetic self-test confirms the rigid-vs-articulated classifier recovers the correct kind for both a straight-line manipulation and a 90° arc at 15 cm radius with centimetre-scale fit RMSE.

TABLE V

LAYER 3 EXTRACTOR ON TEN HOT3D CLIPS. DETECTIONS ARE PER (CLIP, HAND, OBJECT) TUPLE WITH CONTACT DETECTED (15 CM RADIUS, ≥ 10 FRAMES). CLASSIFIER MARGIN: RIGID DETECTIONS SIT AT $+3.1$ MM MEDIAN (*below* THE $+10$ MM THRESHOLD THAT WOULD PUSH THEM INTO THE ARTICULATED BUCKET; THIS IS THE MARGIN BY WHICH THE STRAIGHT-LINE FIT BEAT THE ARC FIT), AND THE TWO ARTICULATED DETECTIONS SIT AT $+14.7$ MM (MARGIN BY WHICH THE ARC FIT BEAT THE STRAIGHT-LINE FIT). WITH ONLY TWO ARTICULATED DETECTIONS WE CANNOT CLAIM THE TWO CLASSES ARE CLEANLY SEPARABLE. DWELL TIMES ARE THE DURATION OF EACH PHASE IN SECONDS (MEAN ACROSS DETECTIONS).

Metric	Rigid	Articulated
detections (n=9 across $6/10$ clips)	7	2
median classifier margin (mm)	$+3.1$	$+14.7$
arc radius (cm, articulated only)	—	7.2
arc swept angle (°, articulated only)	—	169
mean dwell $(t_{\text{app}}, t_{\text{con}}, t_{\text{ret}})$ (s)	$(1.27, 1.20, 1.55)$	

The extractor returns no descriptor on four clips (001849, 002202, 003045, 003219) where the UMETrack wrist either

never comes within 15 cm of any object centroid for 10 consecutive frames, or the wrist holds an object through the entire clip without a distinct approach/retreat. Both are failure modes the extractor should flag; neither is a bug. The $(1.27, 1.20, 1.55)$ s mean dwell is physically reasonable for tabletop grasp-and-move and matches the duration of a single human reach cycle.

E. No-demo ablation (self-match and cross-clip)

The paper's central question is whether the human demonstration adds anything over a hand-coded default. We answer it in two parts.

a) Self-match.: For each clip with a Layer-3 descriptor, plan with the descriptor against the clip's own target object centroid and against the hand-coded default on the same target. Score each plan's RMSE against the clip's GT wrist trajectory over the APPROACH–CONTACT–RETREAT window after arclength resampling.

b) Cross-clip.: For every (S, T) pair of clips with $S \neq T$ and matching descriptor `kind`, apply S's descriptor to T's target object centroid and score against T's GT wrist trajectory. HOT3D object IDs are instance-unique within a clip, so grouping by descriptor `kind` is a looser transfer than class-matched would be. We also fetched the HOT3D object catalog from the public Hugging Face mirror and downloaded additional `train_aria` clips to construct a class-matched cross-clip subset (five classes with ≥ 2 descriptors, 20 pairs total); results for that stricter subset are the bottom block of Table VI.

The two numbers that matter for this submission read in opposite directions and both are now empirically grounded. On self-match, the object-centric v2 planner cuts plan-to-GT RMSE from 19.4 cm (v1) to 13.4 cm and wins $7/9$ of comparisons against the hand-coded default: the demonstration *clearly* carries scene-reconstruction signal, confirmed empirically. On cross-clip transfer, both planners lose to the default (v2 by -6.9 cm, v1 by -3.4 cm, with v2 the worse of the two because its offsets are source-scene-specific). The signal exists; the current planners do not make it scene-portable.

Three readings:

1) **The demo matters when source = target.** The v1 vs v2 gap on self-match ($+6.0$ cm) is a lower bound on how much information the demo descriptor actually encodes. The hand-coded default throws all of that away.
2) **Object-local framing improves pair-level win rate but not the mean.** v4 consumes HOT3D's `T_world_from_object` to express offsets in the demo object's local frame, then rotates them through the target object's local frame at planning time. v1 (world-frame direction) actually has the smallest mean cross-clip loss (-3.42 cm) because it leaves most of the hand-coded default's structure intact; v4 is the configuration we highlight because its win-rate (36% vs v1's 31%) is the one measurable improvement from putting object-local structure into the planner, so it is the strongest demo-driven candidate we tested. The mean plan-to-GT RMSE under v4 is unchanged from v3 (27.36 vs 27.29 cm, both still well above the hand-coded default's 20.85 cm). The class-matched subset (five classes with ≥ 2 descriptors on the HOT3D catalog's 20 pairs) tells a sharper story: aggregate still -6.56 cm, and the one class with a positive mean (*cellphone*, $+1.06$ cm on 6 pairs with $3/6$ wins) is inside coin-flip range for a 6-pair sample. Elongated objects (whiteboard markers and erasers, mean -10 cm, $1/10$ wins) and complex-shape toys (puzzle_toy, -18 cm, $0/2$) lose outright. We do not claim per-class positive transfer from this slice; the reading is that cross-class portability of the current descriptor is not demonstrated.
3) **A class-matched cross-clip test is a separate axis.** Pairing arbitrary rigid descriptors to arbitrary rigid targets ignores object category; a class-matched comparison on the HOT3D catalog (fetched from the public Hugging Face mirror; results in the bottom block of Table VI) reduces the pair-selection noise but is orthogonal to the frame-portability fix.

The reading of the no-demo ablation: the demonstration carries real signal on self-match (v2 cuts plan-to-GT RMSE from 19.4 to 13.4 cm on the clip the demo came from), but neither v3 nor v4 (which were the two planner-side fixes we hypothesised would recover that signal cross-scene) actually beats the hand-coded default on cross-clip; v4 narrowly wins the pair-level race (36% of pairs) but loses the mean by 6.5 cm. The current class-matched subset repeats the loss (-6.56 cm aggregate). We do not have a planner-side fix on offer that we have measured to close the gap.

VI. Discussion

Four findings sit under this submission's results:

(1) Scene-prior composition explains the end-to-end execution numbers. The three protocols that share the hand-coded structured plan on a GT-pose scene (demo, default, oracle) tie on T1 and T2 and are within one trial of each other on T3 (Table I). The gap to the naive wrist baseline measures the contribution of scene-prior-plus-scripted-plan over nothing ($\sim$73 pp on T3); the gap to the no-demo protocol measures the contribution of the Layer-3 demo descriptor over the scripted plan ($\sim$0 pp on T1/T2, $+3.3$ pp strict on T3, inside the CI).

(2) The result generalises to the HOT3D cross-clip ablation (§V-E). On 42 cross-clip HOT3D pairs the demo-derived descriptor loses to the hand-coded default by 6.5 cm plan-to-GT RMSE on average, in the v4 object-oriented frame that was our strongest-hypothesised planner-side fix (Table VI). Pair-level win rate 36% is a slight improvement over v1 but mean loss is unchanged. The 20-pair class-matched subset gives -6.56 cm aggregate; the one subclass with a positive mean (*cellphone* $+1.06$ cm on 6 pairs) is a coin flip.

(3) The "whole descriptor" caveat on T1–T3. The demo protocol in T1–T3 shifts a single waypoint (pre-approach) by a sub-centimetre offset in T1/T2 and $\sim$5 cm in T3. The remaining descriptor fields (approach duration, contact vector, retreat direction, dwell times) are not wired through: our Layer-4 reference planner `plan_from_descriptor_v4` needs

TABLE VI

NO-DEMO ABLATION ACROSS FOUR DESCRIPTOR INSTANTIATIONS: *v1* CARRIES THE DEMO'S WORLD-FRAME APPROACH/RETREAT DIRECTION VERBATIM; *v2* IS OBJECT-CENTRIC (WRIST OFFSETS FROM THE DEMO SCENE'S OBJECT CENTROID); *v3* ADDS A GRAVITY DECOMPOSITION + CAMERA-FORWARD-RELATIVE ROTATION; *v4* USES THE OBJECT'S OWN LOCAL FRAME VIA HOT3D'S `T_WORLD_FROM_OBJECT`, SO A DEMO OFFSET "ABOVE AND SLIGHTLY TO THE RIGHT OF THE OBJECT" TRANSFORMS INTO THE TARGET OBJECT'S OWN ABOVE-AND-RIGHT-OF-HANDLE DIRECTION REGARDLESS OF ITS HEADING. V2 AND V3 COINCIDE IN SELF-MATCH. V4'S CROSS-CLIP WIN RATE NEARLY DOUBLES OVER V3 ($15/42 = 36\%$ VS $8/42 = 19\%$), THE FIRST IMPROVEMENT OVER THE NAIVE V1 WORLD-FRAME BASELINE; THE MEAN RMSE IS UNCHANGED BUT THE STD GROWS (10.0 VS 6.9 CM), SO V4 WINS MORE PAIRS DECISIVELY AND LOSES THE MISMATCHED PAIRS MORE BADLY. THE CLASS-MATCHED SUBSET (BOTTOM BLOCK, 20 PAIRS) DOES *not* FLIP THE MEAN POSITIVE; THE AGGREGATE STAYS AT -6.56 CM AND THE ONE SUBCLASS WITH A POSITIVE SIGN (*cellphone*, $+1.06$ CM ON 6 PAIRS WITH $3/6$ WINS) IS A COIN FLIP RATHER THAN EVIDENCE OF TRANSFER.

Protocol	n	Demo (cm)	Default (cm)	Improvement (cm)
Self-match, v1 (world-frame direction)	9	19.38	21.53	$+2.15$ (6/9 wins)
Self-match, v2 (object-centric offsets)	9	**13.36**	21.53	$+8.17$ (7/9 wins)
Self-match, v3 (gravity-decomposed)	9	**13.36**	21.53	$+8.17$ (7/9 wins)
Self-match, v4 (object-oriented frame)	9	14.34	21.53	$+7.19$ (7/9 wins)
Cross-clip, v1	42	24.26	20.85	-3.42 (13/42 wins)
Cross-clip, v2	42	27.74	20.85	-6.90 (8/42 wins)
Cross-clip, v3	42	27.29	20.85	-6.45 (8/42 wins)
Cross-clip, v4 (object-oriented frame)	42	27.36	20.85	-6.51 ($15/42 \approx 36\%$)
Cross-clip, v4 class-matched (20 pairs)	20	32.64	26.08	-6.56 ($4/20 = 20\%$)
cellphone (6 pairs)	6	—	—	$+\mathbf{1.06}$ ($\mathbf{3/6}$ wins)
keyboard (2 pairs)	2	—	—	-2.33 (0/2)
whiteboard_eraser (6 pairs)	6	—	—	-9.71 (1/6)
whiteboard_marker (4 pairs)	4	—	—	-9.57 (0/4)
puzzle_toy (2 pairs)	2	—	—	-18.15 (0/2)

object-local offsets that the current extractor does not emit (it loads centroids only, not full `T_world_from_object` poses). The T1/T2 null results therefore carry the caveat "null when the demo contribution is a sub-cm nudge"; a whole-descriptor plan might measure differently, and reconciling the extractor-planner interface is the biggest open implementation item for v2 (Section VII).

(4) The cost argument credits the no-demo prior, not the demo. In Table II, the no-demo row has zero demo cost and matches the single-demo pipeline on T1/T2 and is one trial behind on T3. The relevant contrast against VLA fine-tuning rows ($\sim$ 50–200 demos, hours of GPU) is *no-demo*-vs-FT-VLA; our single-demo pipeline adds the 5-second video on top of that at no measurable success-rate gain.

These four findings are consistent with the structural property of T1–T3: top-down pick, same-height place, planar push are default-sufficient tasks. T4 (top-knob mug) tests whether the correct-target oracle can exploit a different grasp feature than the default: paired-McNemar $p=0.074$ strict (not significant at $\alpha = 0.05$) and $p = 0.027$ relaxed (significant); demo and oracle both consume knob position from GT and the gap is bounded by what the default heuristic was tuned to do. T5 (articulated-arc knob turn) is the paper's one clear positive demo-vs-default result: the task structurally requires sweeping the EE along an arc, the extracted Layer-3 arc-parameters make that possible, and the no-descriptor default cannot even approach the gate—paired-McNemar $p = 0.0026$ strict and $p=3.6\times10^{-5}$ relaxed. T5 does *not* close the extraction-from-video question (its demo column consumes the extractor's output but the extractor's input is GT UMETrack mocap, not a monocular pipeline); closing that gap—wiring MediaPipe + MiDaS + COLMAP into the Layer-3 extractor—is the biggest open implementation item (Section VII).

VII. LIMITATIONS

Not end-to-end monocular, and the fine-tuned Layer 1 *does not fully close the gap.* Layer-3 descriptors are extracted from HOT3D's GT UMETrack hand/object poses, not from the deployed Layer-1/2 perception stack. The off-the-shelf SE(3)-only Layer-1 test (§V-A) measures 47.87 cm best-of-five mean wrist RMSE across all backbones we tried. The Layer-3 articulated classifier fits an arc with a $\sim$1 cm ($\sim$10 mm) RMSE threshold on wrist landmarks, so wrist noise much above that level is expected to dominate the extracted arc parameters. Our HOT3D fine-tune of the DAV2 Metric decoder brings held-out wrist RMSE to 13.71 cm pooled / 4.07 cm pooled median on 4 test clips—a 4–9$\times$ improvement that still sits several times above the Layer-3 margin; the fine-tune narrows the gap substantially in-distribution but does not fully close it. Whether the residual $\sim$4 cm leaves enough headroom for the classifier to recover the 7.2 cm arc radius inside its margin is the open empirical question. Wiring the fine-tuned Layer-1 chain into the Layer-3 extractor and reporting T5 on an extraction that consumes monocular video is the natural next experiment and the biggest open scoping limitation of the current submission.

Whole-descriptor plan rewiring. The demo protocol in T1–T3 shifts only the pre-approach waypoint by sub-cm (T1/T2) or $\sim$ 5 cm (T3); the rest of the Layer-3 descriptor (approach duration, contact vector, retreat direction, dwell times) does not flow into the plan. Our Layer-4 reference planner `plan_from_descriptor_v4` expects object-local offsets that the current extractor does not emit (centroids only, not full `T_world_from_object` poses). Reconciling the extractor-planner interface is the biggest open implementation item for v2; the T1/T2 null carries the caveat "null when the demo contribution is a sub-cm nudge".

Default-sufficient control group. T1–T3 are a control group

for "when the demo should *not* matter": top-down pick, same-height place, planar push are exactly what a straight-down scripted plan handles. T4 and T5 are the structural tasks: T4 (top-knob mug) is where the default fails because the grasp feature differs between protocols, and T5 (articulated-arc knob turn) is where the default fails because the required EE motion is a swept arc the default cannot produce. Extending the structural family—pouring (orientation-sensitive), hinged-door opening on a hinge whose geometry is *randomised independently of the demo* (the current T5 hinge is placed to coincide with the extracted frame), wiping along a path, angled insertion—and pairing each with a real end-to-end Layer-3 extraction is the natural next-paper experiment. Reachability on side-approach grasps depends on the arm's DoF count; the 7-DoF MuJoCo Franka Panda used here handles all of the above, a 5-DoF arm (planned for the real-robot row in a follow-up) would not.

T5 scene-descriptor coupling. The T5 sim scene is built to match the extracted descriptor: the lever length is set to the extracted 7.2 cm radius and the hinge is placed so the extracted arc frame coincides with the scene frame. This design is why *demo* and *correct-target oracle* coincide in T5 (both consume the same arc) and why we can isolate the planner's ability to execute the arc; it also means T5 cannot distinguish "the extractor measured the arc correctly" from "the scene was built around the extraction." A follow-up test that randomises the hinge geometry *independently* of the descriptor (lever length drawn from, e.g., 5–10 cm; hinge location perturbed) would turn the *oracle* column into a measurement of extraction fidelity against GT hinge geometry.

T5 default is a lift-up straw-man. The no-demo default in T5 grips the knob and lifts straight up; it has no articulation prior at all. A compliant "grasp and let the joint guide the motion" baseline (analogous to a door-handle heuristic) would be a stronger default and would narrow the demo-vs-default gap. We did not run that baseline here; the honest reading of the T5 p-values is "significantly better than a default with no articulation prior," not "significantly better than every reasonable default."

T5 demonstration source is a wrist arc around a rigid object. HOT3D clip-002062 object 24 is a rigid object per the dataset annotation; our Layer-3 articulated classifier fits an arc to the wrist trajectory as the user reorients the object in hand. T5 therefore tests whether an arc extracted from a human reorientation motion can drive a sim knob-turn where the robot's EE sweeps an arc; the human and the robot are not performing the same manipulation. This is a defensible trajectory-shape transfer, not an articulated-object transfer.

Short-video coverage. HOT3D clips are 300-frame segments ($\sim$5 s); multi-minute YouTube clips are out of scope for the current segmentation.

Scene prior assumes a single horizontal counter. Multi-plane or non-horizontal workspaces are out of scope for the current Layer 2 RANSAC fitter.

Simulator-only execution. All numbers are MuJoCo; a real-robot row is strengthening-tier.

VIII. Conclusion

A controlled measurement under matched scene scaffolding finds that a Layer-3 affordance descriptor extracted from one 5-second human-video clip does not measurably change execution success over a no-demo hand-coded planner on three default-sufficient MuJoCo Franka tasks (T1/T2 paired-McNemar $p = 1$ with $b = c = 0$; T3 exact McNemar $b = 1$, $c = 0$, $p = 1$). The result is corroborated by a 42-pair cross-clip ablation (§V-E) (-6.5 cm plan-to-GT RMSE for the v4 object-oriented descriptor on average, 36% win rate; the class-matched subset confirms it at -6.56 cm). The scene-prior composition is what works on these task shapes; the per-clip affordance descriptor adds nothing measurable there. Two structural tasks exit the default-sufficient regime: T4 (top-knob mug) at 16.7% | 36.7% oracle/demo vs 0% | 10% no-demo default (paired-McNemar $p = 0.074$ strict, $p = 0.027$ relaxed; demo = oracle, both consume knob position from GT), and T5 (articulated-arc knob turn) at **36.7**% | **63.3**% oracle/demo vs 0% | 0% no-demo default (paired-McNemar $p = 0.0026$ strict, $p = 3.6\times10^{-5}$ relaxed). T5 is the submission's one clear positive demo-vs-default result: the task structurally requires sweeping the EE along a circular arc of the Layer-3-extracted radius, the arc-aware planner succeeds, and the no-descriptor default cannot even approach the gate. The pipeline is not yet end-to-end monocular in the execution experiments—Layer-3 consumes GT UMETrack mocap, not Layer-1/2 output—but we narrow the Layer-1 gap substantially in isolation: a $\sim$5-min fine-tune of DAV2 Metric Indoor's decoder on sparse wrist supervision across 14 HOT3D train clips brings held-out wrist RMSE from 128.81 cm out-of-box to 13.71 cm pooled / 4.07 cm pooled median on 4 never-seen test clips (§V-A)—a 4–9$\times$ in-distribution reduction, with the residual $\sim$4 cm still above the Layer-3 articulated classifier's $\sim$10 mm arc-fit margin. Wiring the fine-tuned Layer-1 into the Layer-3 extractor and rerunning T5 on a video-driven descriptor is the natural next step and the test of whether the residual 4 cm error is small enough for the arc parameters to survive extraction.

References

[1] Anonymous, "F.R.A.N.T.Z.: 3D scene reconstruction and hand-trajectory extraction from monocular video via training-light compositional AI," *Anonymised preprint*, 2026, author list and venue redacted for double-blind review; arXiv ID to be inserted on companion submission.

[2] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[3] Octo Model Team, "Octo: An open-source generalist robot policy," *arXiv preprint arXiv:2405.12213*, 2024.

[4] M. J. Kim, K. Pertsch, S. Karamcheti *et al.*, "OpenVLA: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[5] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak, "Affordances from human videos as a versatile representation for robotics," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[6] S. Bahl, A. Gupta, and D. Pathak, "Human-to-robot imitation in the wild (WHIRL)," in *Robotics: Science and Systems (RSS)*, 2022.

[7] H. Bharadhwaj, R. Mottaghi, A. Gupta, and S. Tulsiani, "Track2act: Predicting point tracks from internet videos enables diverse zero-shot robot manipulation," in *European Conference on Computer Vision (ECCV)*, 2024.

[8] C. Wen, X. Lin, H. Che, C. Bai, and D. Jayaraman, "Any-point trajectory modeling for policy learning (ATM)," in *Robotics: Science and Systems (RSS)*, 2024.

[9] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, and Y. Zhu, "MimicPlay: Long-horizon imitation learning by watching human play," in *Conference on Robot Learning (CoRL)*, 2023.

[10] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu, "EgoMimic: Scaling imitation learning via egocentric video," *arXiv preprint*, 2024.

[11] J. Li, Y. Zhu, Y. Xie, Z. Jiang, M. Seo, G. Pavlakos, and Y. Zhu, "OKAMI: Teaching humanoid robots manipulation skills through single video imitation," in *Conference on Robot Learning (CoRL)*, 2024.

[12] M. Lepert, J. Fang, and J. Bohg, "Phantom: Training robots without robots using only human videos," *arXiv preprint*, 2025.

[13] Z. Liu, D. Niu, B. Zhang, B. Shao, J. Wang, R. Zheng, Y. Sharma, R. Lu, J. Liu, R. Zhang, R. Herzig, Y. Bai, K. Goldberg, D. Xu, Y. Zhu, G. Wang, L. Fan, and T. Darrell, "Agentic world action models," *arXiv preprint arXiv:2610.30136*, 2026.

[14] A. Brohan *et al.*, "RT-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[15] Physical Intelligence, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[16] S. Tao, F. Xiang, A. Shukla, Y. Qin, X. Hinrichsen, X. Yuan, C. Bao, X. Lin, Y. Liu, T. kai Chan, Y. Gao, X. Li, T. Mu, N. Xiao, A. Gurha, Z. Huang, R. Calandra, R. Chen, S. Luo, and H. Su, "ManiSkill3: GPU parallelized robotics simulation and rendering for generalizable embodied AI," in *Robotics: Science and Systems (RSS)*, 2025.

[17] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012, pp. 5026–5033.

[18] C. C. Gordon, C. L. Blackwell, B. Bradtmiller, J. L. Parham, P. Barrientos, S. P. Paquette, B. D. Corner, J. M. Carson, J. C. Venezia, B. M. Rockwell, M. Mucher, and S. Kristensen, "2012 Anthropometric Survey of U.S. Army Personnel: Methods and Summary Statistics," U.S. Army Natick Soldier Research, Development and Engineering Center, Tech. Rep. NATICK/TR-15/007, 2014.

[19] L. Piccinelli, Y.-H. Yang, C. Sakaridis, M. Segu, S. Li, L. Van Gool, and F. Yu, "UniDepth: Universal monocular metric depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[20] W. Yin, C. Zhang, H. Chen, Z. Cai, G. Yu, K. Wang, X. Chen, and C. Shen, "Metric3D: Towards zero-shot metric 3D prediction from a single image," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[21] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, K. Wang, G. Yu, C. Shen, and S. Shen, "Metric3Dv2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.