LEARNING PHYSICALLY GROUNDED LATENT ACTIONS FROM HUMAN AND ROBOT VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent-action pretraining provides a scalable way to learn robot policies from actionless videos, but existing objectives are driven primarily by visual change and therefore provide only indirect supervision of physical interaction. Tactile and force measurements can provide direct physical signals, but they are available for only a limited subset of existing human and robot videos. Our key insight is to decouple the acquisition of physical supervision from its transfer to latent-action learning: physical supervision derived from a limited corpus of instrumented interactions can be propagated to broader collections of human and robot videos without paired sensor measurements. Building on this insight, we propose **PHYSLA**, a **Phys**ically grounded **L**atent-**A**ction framework that learns a shared, horizon-structured representation from heterogeneous physical measurements, including pressure distributions, tactile marker displacements, and joint torques, and infers these representations from causal RGB observations. During vision-language-action pretraining, the inferred physical targets and future visual prediction jointly supervise the same latent actions. Importantly, the transferred targets capture interaction-relevant latent structure rather than calibrated sensor readings, allowing physical supervision to scale beyond the instrumented corpus while requiring no tactile or torque sensing at deployment. **PHYSLA** achieves 98.2% average success on LIBERO, 74.0% and 64.7% on SimplerEnv's Google Robot and WidowX settings, respectively, and 65.5% across five real-world G1 tasks. These results support the effectiveness of transferring measurement-derived physical supervision to latent-action learning from human and robot videos.

1 INTRODUCTION

Scaling vision-language-action (VLA) models requires diverse interaction experience from robot and human demonstrations (Ma et al., 2026). Robot trajectories provide embodiment-aligned observations and action supervision but remain costly to collect at scale (Open X-Embodiment Collaboration, 2024; Khazatsky et al., 2024), whereas human videos offer broader manipulation diversity without direct supervision in robot action space (Grauman et al., 2022; Hoque et al., 2025; Punamiya et al., 2026). Latent-action learning incorporates the existing actionless videos into policy pretraining by encoding video transitions. LAPA (Ye et al., 2024) learns discrete latent codes through future-frame reconstruction. UniVLA (Bu et al., 2025) further uses language conditioning to suppress task-irrelevant dynamics, while VLA-JEPA (Sun et al., 2026) supervises latent actions via future visual representation prediction. Nonetheless, physical measurements are paired with only a limited subset of videos, and objectives grounded primarily in visual dynamics may encode camera motion artifacts and manipulation-irrelevant changes, leaving physical interaction insufficiently represented (Zhang et al., 2025).

Recent methods further strengthen latent-action supervision through optical-flow consistency (Routray et al., 2025), future robot state and action prediction (Chen et al., 2025a), or alignment with robot action vocabularies (Zhang et al., 2026b). These signals characterize contact and mechanical response only indirectly: similar visible movements can correspond to different contact conditions, while pressure or loading may change under subtle displacement. Physical measurements therefore provide complementary supervision that is difficult to recover from motion or action alignment alone. However, the scalability of physical supervision is constrained by the limited availability of paired pressure, tactile, and torque measurements in human and robot video datasets.

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

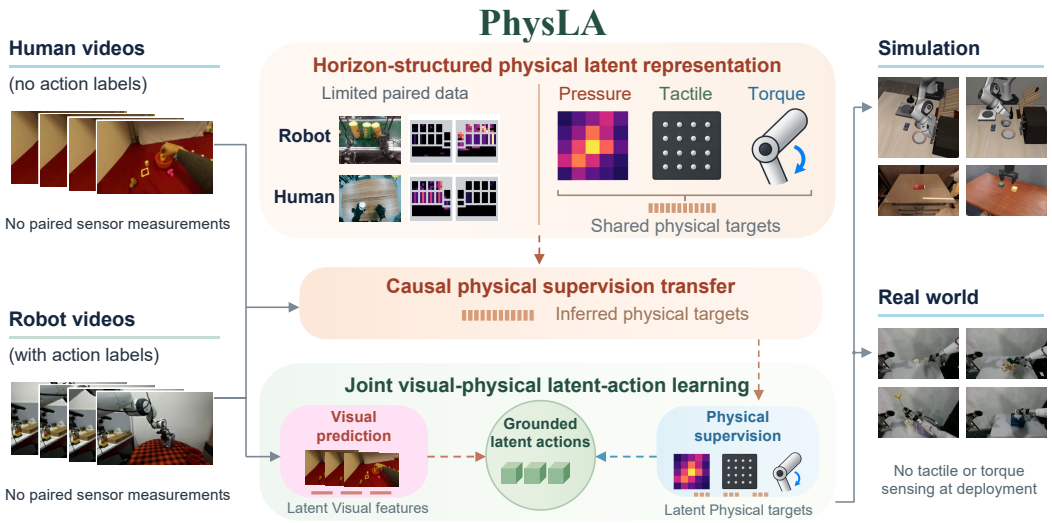

Figure 1: **Scaling physical supervision beyond instrumented data. PHYSLA** learns temporally structured physical targets from limited paired RGB-sensor data and infers them from causal RGB for uninstrumented human and robot videos. The inferred targets complement visual prediction during latent-action pretraining, while deployment requires no physical sensing.

This raises the central question of this work: *how can physical supervision acquired from a limited corpus of instrumented interactions be transferred to latent-action learning from human and robot videos without paired sensor measurements?*

Our key insight is to transfer physical supervision through a shared representation learned from instrumented interactions and inferred from causal RGB observations. As illustrated in Figure 1, the inferred representations provide physical targets for latent-action learning on broader collections of human and robot videos without paired sensor measurements. In this work, *physical grounding* refers to supervising latent actions with representations learned from heterogeneous physical measurements across human and robot interactions, including pressure distributions, tactile marker displacements, and joint torques, which provide complementary physical signals of contact, deformation, and mechanical response.

Building on this insight, we introduce **PHYSLA**, a framework that learns horizon-structured targets from heterogeneous physical measurements and infers them from causal RGB history to supervise latent actions on human and robot videos. These targets provide physical supervision signals and complement future visual prediction during VLA pretraining, while robot trajectories additionally provide action supervision. The resulting policy benefits from physical grounding without requiring physical sensors or target generators at deployment. **PHYSLA** achieves average success rates of 98.2% on LIBERO and 65.5% across five real-world manipulation tasks, compared with 97.2% and 49.8% for VLA-JEPA, respectively.

Our contributions are threefold:

- We introduce a paradigm for transferring physical supervision from limited instrumented interactions to human and robot videos without paired sensor measurements. It learns shared, horizon-structured representations from heterogeneous physical measurements and infers them from causal RGB history.
- We propose **PHYSLA**, a unified latent-action pretraining framework in which future visual prediction and transferred physical targets jointly supervise the same latent actions across human and robot streams. The resulting policy benefits from physical grounding without requiring physical sensor inputs or target-generation modules at deployment.
- We assess physical representation quality, transfer through causal RGB inference, and downstream manipulation. Controlled ablations examine the contribution of physical supervision and the effect of applying it to human videos and robot videos.

2 Related Work

2.1 Latent Action Learning and Latent Predictive Modeling

Latent-action pretraining learns intermediate action representations from video transitions, supporting policy learning from actionless videos. LAPA (Ye et al., 2024) learns discrete latent codes through future-frame reconstruction, while Moto (Chen et al., 2025b) introduces latent motion tokens for pretraining. Reconstruction can also encode incidental visual changes (Zhang et al., 2025). UniVLA (Bu et al., 2025) uses language conditioning to suppress task-irrelevant dynamics, and VLA-JEPA (Sun et al., 2026) supervises latent actions through future visual representation prediction. Additional supervision connects latent actions to motion and robot behavior. ViPRA (Routray et al., 2025) uses perceptual and optical-flow consistency, villa-X (Chen et al., 2025a) predicts robot states and actions, and CLAP (Zhang et al., 2026b) aligns visual transitions with a robot action vocabulary. Our work instead adds supervision derived from physical measurements and transfers it through causal RGB inference to latent-action learning on both human and robot videos.

2.2 Latent Predictive Modeling for Robot Learning

Latent dynamics support planning and representation learning. TD-MPC2 (Hansen et al., 2024), PLDM (Sobal et al., 2025), and DINO-WM (Zhou et al., 2024) use learned latent dynamics for control and planning. DynaMo (Cui et al., 2024) learns latent inverse and forward dynamics without ground-truth actions, while V-JEPA 2 (Assran et al., 2025) combines action-free video pretraining with action-conditioned robot post-training. JEPA-VLA (Miao et al., 2026) integrates predictive visual embeddings into policies; VLA-JEPA (Sun et al., 2026) uses future feature prediction to supervise latent actions. We draw on predictive learning rather than introduce a standalone world model. Future visual prediction and horizon-structured physical targets jointly supervise the same latent actions. Inferred from causal RGB, these measurement-derived targets extend physical supervision to human and robot videos without paired sensor recordings.

2.3 Physical Interaction Supervision for Robot Learning

Physical sensing reveals interaction information that vision may not fully observe. Visuo-tactile policies use touch for manipulation (Calandra et al., 2018); Sparsh (Higuera et al., 2024), T3 (Zhao et al., 2024), and AnyTouch (Feng et al., 2025) learn transferable tactile features. Paired observations also support tactile prediction and visual-tactile alignment (Takahashi & Tan, 2019; Yang et al., 2022). EgoTac (Zhang et al., 2026c) learns from paired visual–tactile data to predict force and contact from egocentric videos without tactile inputs at inference, while TTP (Zhang et al., 2026a) uses future tactile prediction on instrumented human demonstrations for human-to-robot skill transfer. In contrast, we transfer shared latent targets from heterogeneous physical measurements to human and robot videos through causal RGB inference, jointly supervising latent actions with visual prediction. These targets encode interaction-relevant structure rather than calibrated sensor readings, and deployment requires no physical sensing.

3 Method

3.1 Physically Grounded Latent Action Framework

We connect physical representation learning with policy pretraining through a shared space of ordered supervision targets. Figure 2 summarizes three stages: physical representation learning, causal supervision transfer, and joint visual–physical pretraining. Let x_t denote the policy context, c_t^v the RGB history for target inference, and p_t^+ the aligned future physical measurements. Both contexts are causal at t, but c_t^v may span a longer history. Future measurements define training targets and are never policy inputs.

Physical representation learning. Masked reconstruction trains a shared encoder to map heterogeneous measurements to ordered targets, $y_t^* = E_{\text{phys}}(p_t^+) = [y_t^{*,1}, \ldots, y_t^{*,M}]$ (Section 3.2). Each target is associated with a temporal region, giving subsequent stages a common supervision format across measurement sources.

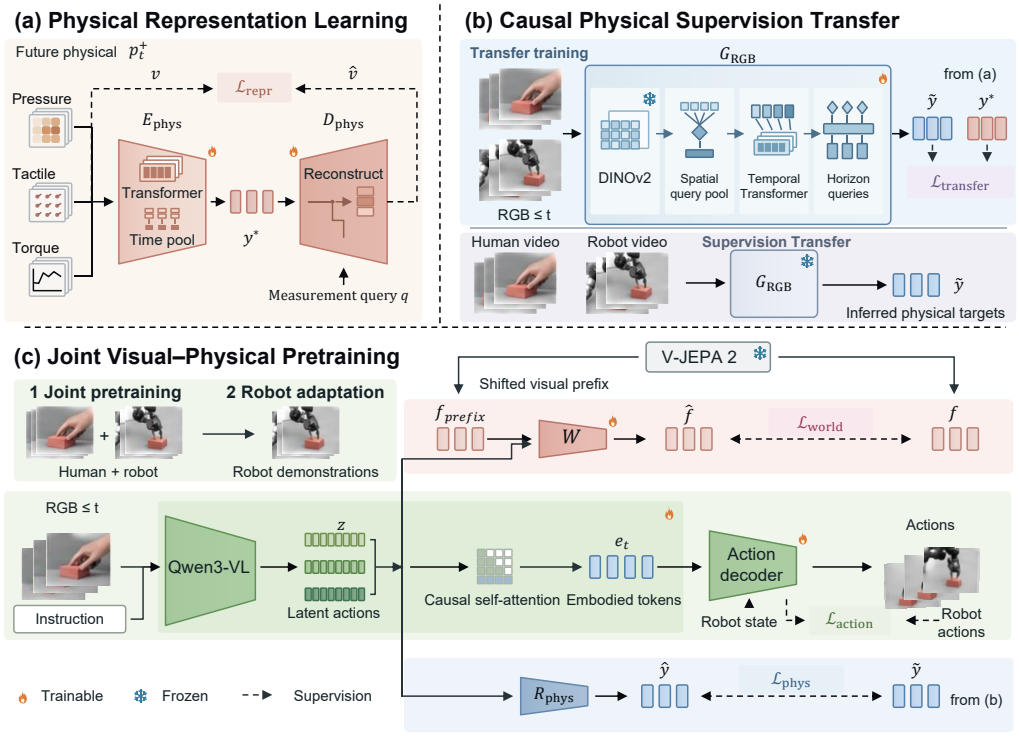

Figure 2: **PHYSLA overview.** (a) Masked reconstruction learns ordered physical targets. (b) Causal RGB inference transfers these targets to human and robot videos. (c) Visual and physical objectives supervise shared latent actions; robot trajectories additionally supervise action generation. Target generators remain frozen during joint pretraining.

Transfer of physical supervision. With E_{phys} frozen, paired RGB and measurements train G_{RGB} to infer $\tilde{y}_t = G_{\text{RGB}}(c_t^v)$, providing physical targets for additional human and robot videos (Section 3.3). Thus, y_t^* is extracted from measurements, whereas $\tilde{y}_t$ estimates that representation from visual evidence.

Joint latent-action learning. The VLA backbone produces latent actions z_t from x_t. A physical prediction head produces $\hat{y}_t = R_{\text{phys}}(z_t)$, aligned with valid transferred targets $\tilde{y}_t$. Physical alignment and future visual prediction supervise the same latent actions in both streams; robot trajectories additionally provide action supervision (Section 3.4). The target generators remain frozen, and physical predictions do not provide inputs to the action decoder.

3.2 HORIZON-STRUCTURED PHYSICAL LATENT REPRESENTATION

Physical interaction evolves within a prediction window through changes in contact, loading, and release. A single window-level target may obscure these temporal distinctions. We therefore encode future measurements into ordered, horizon-specific targets. To accommodate differences in sensor layout and dimensionality, each measurement j uses a common token interface:

$$u_j = \phi\big([v_j, \Delta v_j, m_j^v, s_j, m_j^s, \tau_j]\big)\,, \tag{1}$$

where v_j, Δv_j, s_j, and τ_j are the normalized value, temporal difference, sensor-local coordinate, and relative time. Masks m_j^v and m_j^s indicate valid numerical and coordinate components. This interface accommodates pressure grids, tactile marker displacements, and joint torques without dataset or sensor-identity tokens. Source-specific training statistics normalize measurements before differencing; invalid tokens and padding are excluded from attention.

A shared Transformer contextualizes tokens across the full window. We partition time into M consecutive, non-overlapping regions $\{\mathcal{I}_h\}_{h=1}^{M}$, pool encoded tokens by their original timestamps,

and apply a shared projection:

$$\{r_j\} = T_{\text{phys}}(\{u_j\}), \qquad y_t^{*,h} = P_{\text{phys}}\left(\frac{1}{\max(1, |\mathcal{V}_h|)} \sum_{j \in \mathcal{V}_h} r_j\right), \tag{2}$$

where $\mathcal{V}_h$ contains valid, unmasked tokens in $\mathcal{I}_h$ and $y_t^{*,h} \in \mathbb{R}^D$. Contextualization precedes pooling: each target retains a distinct temporal assignment while incorporating information from the full window. Sampling and dimensions appear in Appendix A.

Masked reconstruction trains the horizon targets to retain measurement information. Masked values and their dependent temporal-difference features are excluded from encoder inputs to prevent reconstruction through direct access to those values. Decoder D_{phys} reconstructs each masked value from its horizon target and query q_j, which encodes the sensor-local coordinate, coordinate-validity mask, and relative time:

$$\mathcal{L}_{\text{repr}} = \frac{1}{|\Omega|} \sum_{(j,k) \in \Omega} \rho\Big([D_{\text{phys}}(y_t^{*,h(j)}, q_j)]_k - v_{j,k}\Big), \tag{3}$$

Here, Ω indexes masked, valid numerical components, $h(j)$ identifies the temporal region, and ρ is SmoothL1. After training, the decoder is discarded and E_{phys} is frozen; targets are extracted in evaluation mode from all valid measurements with random masking disabled.

3.3 Video-Conditioned Physical Supervision Transfer

With E_{phys} frozen, paired video and measurements train G_{RGB} to infer ordered targets $\widetilde{y}_t = [\widetilde{y}_t^1, \ldots, \widetilde{y}_t^M]$ from RGB history c_t^v ending at t. Freezing E_{phys} fixes the target space while G_{RGB} learns to infer it from visual evidence. Physical measurements define the supervision; all RGB inputs are available at or before t.

Frozen DINOv2 (Oquab et al., 2023) extracts patch features, which learned spatial queries aggregate into frame tokens. A temporal Transformer with positional embeddings encodes the history. One query per horizon attends to this history, and a shared projection produces M targets in $\mathbb{R}^D$. Horizon-wise alignment preserves the correspondence to the temporal regions defined by E_{phys}. Training uses feature-normalized SmoothL1:

$$\mathcal{L}_{\text{transfer}} = \frac{1}{M} \sum_{h=1}^{M} d_h\Big(\widetilde{y}_t^h, y_t^{*,h}\Big), \quad d_h(a,b) = \frac{1}{D} \sum_{k=1}^{D} \rho\left(\frac{a_k - b_k}{\sigma_{h,k}^{\text{eff}}}\right), \tag{4}$$

where $\sigma_{h,k}^{\text{eff}} = \max(\sigma_{h,k}, 10^{-4})$ and $\sigma_{h,k}$ is the target standard deviation computed on the paired training split for horizon h and feature k. This scaling balances the contributions of target dimensions with different variances. After training, G_{RGB} remains frozen and supplies the same target format to human and robot videos without paired measurements. These inferred representations provide supervision for joint latent-action learning in Section 3.4.

3.4 Joint World-Physical Latent Action Learning

We adopt VLA-JEPA's latent-action and embodied-action architecture (Sun et al., 2026), adding transferred physical targets to its visual prediction objective. Both objectives act on the same latent actions, connecting visual transition modeling with measurement-derived interaction supervision.

Joint policy pretraining. Human and robot batches alternately update a shared Qwen3-VL-2B-Instruct backbone (Bai et al., 2025). Given causal context x_t, it produces horizon-grouped latent actions z_t followed by embodied-action tokens e_t. Through causal attention, e_t attends to the preceding latent actions and, together with robot state s_t, conditions the action decoder. Robot action supervision thereby connects the shared video representation to executable control. A world predictor W predicts future visual features from frozen V-JEPA 2 ViT-L/16 (Assran et al., 2025) using z_t, minimizing mean absolute error $\mathcal{L}_{\text{world}}$ across horizons and feature dimensions. The target visual features supervise prediction and are not policy inputs. Model configurations appear in Appendix A.

A temporal physical prediction head models dependencies across horizons and produces $\widehat{y}_t = R_{\text{phys}}(z_t) = [\widehat{y}_t^1, \ldots, \widehat{y}_t^M]$. It aggregates each horizon's latent slots with a summary token and

attention restricted to the current and earlier horizon groups. The predictions are aligned with the inferred targets through

$$\mathcal{L}_{\text{phys}} = \frac{1}{M} \sum_{h=1}^{M} \ell_{\text{phys}}\big(\widehat{y}_t^h, \widetilde{y}_t^h\big)\,, \tag{5}$$

where ℓ_{phys} is dimension-averaged SmoothL1 on unstandardized targets. Visual and physical targets correspond by temporal order, without requiring identical temporal support. Human batches optimize world and physical losses; robot batches additionally optimize the flow-matching action loss:

$$\begin{aligned} \mathcal{L}_{\text{human}} &= \lambda_{\text{world}}\mathcal{L}_{\text{world}} + \lambda_{\text{phys}}\mathcal{L}_{\text{phys}}, \\ \mathcal{L}_{\text{robot}} &= \lambda_{\text{act}}\mathcal{L}_{\text{action}} + \lambda^r_{\text{world}}\mathcal{L}_{\text{world}} + \lambda_{\text{phys}}\mathcal{L}_{\text{phys}}. \end{aligned} \tag{6}$$

We use $\lambda_{\text{act}} = \lambda_{\text{world}} = 1$ and $\lambda^r_{\text{world}} = \lambda_{\text{phys}} = 0.1$. Physical loss applies only to valid aligned targets in either stream. Its gradients update R_{phys} and the shared backbone, while E_{phys} and G_{RGB} remain frozen.

Downstream adaptation. For LIBERO, SimplerEnv, and G1, we adapt Full using action-labeled robot demonstrations while retaining the action, world, and physical objectives. We update the shared backbone, world predictor, and action decoder while freezing R_{phys}. The world predictor and action decoder also update. Benchmark-specific settings appear in Appendix D. Deployment uses only vision, language, and robot state, without target-generation or auxiliary prediction modules.

4 Experiments

We conduct our experiments around three questions: *(i)* whether physical representations capture temporally structured interaction information, *(ii)* whether causal RGB can recover this information for videos without physical measurements, and *(iii)* whether the transferred supervision improves downstream manipulation. Masked reconstruction, temporal controls, and frozen-feature probes evaluate representation quality and transfer. Simulation and real-world benchmarks assess downstream performance, while source ablations isolate the effects of physical supervision on human videos, robot videos, and their combination.

4.1 Experimental Setup

Pretraining data. To learn and transfer the physical representation, we use synchronized RGB-sensor sequences from EgoTouch (Zhou et al., 2026), comprising pressure grids; VLA-Touch (Bi et al., 2026), comprising tactile marker displacements; and Franka trajectories from RH20T (Fang et al., 2024), comprising joint torques. During VLA pretraining, robot batches are drawn from DROID (Khazatsky et al., 2024), while action-unlabeled video batches are drawn from EgoDex (Hoque et al., 2025), EgoTouch (Zhou et al., 2026), and Something-Something V2 (SSV2) (Goyal et al., 2017). Future visual prediction supervises both streams, while physical alignment is applied to samples with valid transferred targets. Robot action supervision is available only for DROID. Dataset statistics, splits, and training details are provided in Appendix A.

Baselines and controls. For downstream control, we compare **PhysLA** with general-purpose VLA policies, including π_0, π_0-fast, $\pi_{0.5}$, OpenVLA-OFT, and GR00T N1 (Black et al., 2024; Pertsch et al., 2025; Physical Intelligence et al., 2025; Kim et al., 2025; NVIDIA, 2025), as well as video-based latent-action methods, including LAPA, Moto, UniVLA, villa-X, and VLA-JEPA (Ye et al., 2024; Chen et al., 2025b; Bu et al., 2025; Chen et al., 2025a; Sun et al., 2026). For representation evaluation, we compare with DINOv2 (Oquab et al., 2023), TouchAnything's pressure-supervised prehead (PS-Prehead) (Zhou et al., 2026). The physical encoder E_{phys} directly observes measurements and serves as a privileged reference.

Evaluation protocol. We evaluate the physical latent space using masked-reconstruction error and frozen linear probes, and assess physical information in RGB-derived representations and policy latent actions using independently fitted linear probes. We report R^2 for continuous physical attributes, AUROC for binary contact events, and task success rates for downstream control. The LIBERO source ablations share initialization, training data, optimization and adaptation budgets, and rollout configurations; physical-loss routing follows Appendix C. Benchmark-specific protocols appear in Appendix D.

Table 1: **Physical information in video representations and latent actions.** Three-horizon means: R^2 for continuous attributes and AUROC for events. (↑). PS-Prehead also uses hand pose.

Representation	Split	Pressure	Change	|Change|	Contact	Onset	Release
(a) Causal video representations							
DINOv2	Seen	0.4303	-0.0240	0.1774	0.7874	0.7402	0.7325
PS-Prehead	Seen	0.5107	0.0128	0.1934	0.9227	0.8190	0.7924
G_{RGB} (ours)	Seen	**0.6730**	**0.3267**	**0.3322**	**0.9465**	**0.8738**	**0.8898**
DINOv2	Unseen	0.1886	-0.0302	0.0587	0.7974	0.6873	0.6989
PS-Prehead	Unseen	0.4779	0.0141	0.0949	0.8837	0.7974	0.8139
G_{RGB} (ours)	Unseen	**0.5348**	**0.1539**	**0.1728**	**0.9622**	**0.9077**	**0.9064**
(b) Frozen policy latent-action tokens							
w/o physical	Seen	0.4077	-0.0371	0.1806	0.7596	0.7392	0.7205
PHYSLA (ours)	Seen	**0.6335**	**0.3649**	**0.4854**	**0.8715**	**0.8967**	**0.9124**
w/o physical	Unseen	0.0857	-0.0652	-0.0055	0.7194	0.6399	0.6380
PHYSLA (ours)	Unseen	**0.3772**	**0.2100**	**0.2816**	**0.8921**	**0.8189**	**0.8355**

4.2 PHYSICAL REPRESENTATION EVALUATION

We evaluate E_{phys} using masked reconstruction and frozen-feature probes. The temporal controls preserve the original reconstruction targets while either reflecting time coordinates or permuting horizon targets with fixed decoder queries (see Appendix B.3). Ordered targets yield lower reconstruction error than the zero reference and both temporal controls. Horizon permutation causes the largest increase, indicating that reconstruction depends on correct target-to-time correspondence.

As shown in Table 8 and Table 9, frozen linear probes accurately recover pressure levels, signed pressure changes, and robot measurements from E_{phys}, although pressure-change magnitude is recovered less accurately. E_{phys} directly encodes physical measurements, it serves as a privileged reference rather than a visual baseline.

4.3 EVALUATION OF VIDEO-TO-PHYSICAL REPRESENTATION TRANSFER

Table 1(a) evaluates frozen video representations using independently fitted linear probes. G_{RGB} outperforms DINOv2 and TouchAnything's pressure-supervised representation (PS-Prehead) on all six attributes across both splits. DINOv2 uses the same causal RGB history, whereas PS-Prehead additionally uses hand pose.

For G_{RGB}, pressure-level and temporal-change R^2 decrease on unseen tasks, while contact-event AUROC remains high. Contact-event information therefore transfers more consistently across the task splits than pressure magnitude and change. G_{RGB} also outperforms DINOv2 on robot measurements, achieving positive R^2 on all reported torque and marker-displacement attributes (Table 9). Mean-torque R^2 is 0.3264 versus -0.1393, while marker-displacement-change R^2 is 0.2747 versus -0.0349. The remaining gap to E_{phys} indicates that RGB recovers physical information only partially. Figure 3 illustrates this partial recovery during pillow arrangement: the PhysLA-based pressure and active-area estimates track the measured changes more closely than the DINOv2 baseline in this excerpt, although discrepancies in magnitude remain.

4.4 SIMULATION EXPERIMENTS

Our policy achieves an average success rate of 98.2% across the four LIBERO suites (Liu et al., 2023), compared with the VLA-JEPA result of 97.2% (Table 2). It improves on Spatial, Goal, and Long while matching 99.6% on Object. Under the SimplerEnv visual-matching protocol (**?**), our policy achieves 74.0% on Google Robot and 64.7% on WidowX, exceeding VLA-JEPA averages of 65.2% and 57.3%, respectively (Table 12). It scores higher on all eight task groups, although performance varies across tasks.

4.5 REAL-WORLD ROBOT EXPERIMENTS

Across five G1 manipulation tasks, our policy achieves an average success rate of **65.5%**, compared with 49.8% for VLA-JEPA and 49.3% for $\pi_{0.5}$ (Table 3). It achieves higher success on every task,

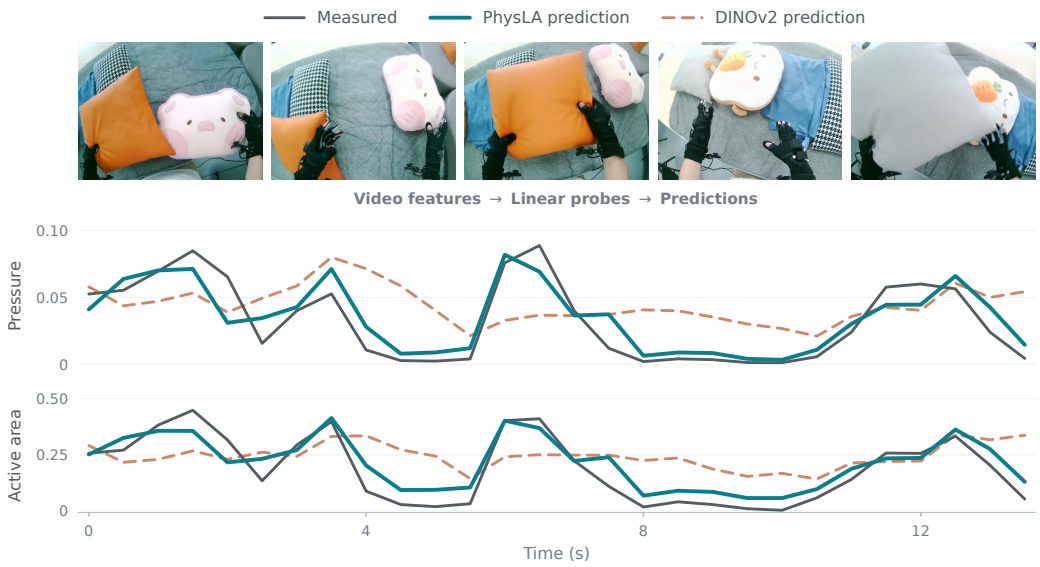

Figure 3: **Physical-attribute decoding on EgoTouch.**

Independent linear probes decode pressure and active sensor area from frozen video representations during pillow arrangement.

Table 2: LIBERO success rates (%). Long denotes LIBERO-10.

Method	Spatial	Object	Goal	Long	Avg.
π_0 (Black et al., 2024)	96.8	98.8	95.8	85.2	94.2
π_0-fast (Pertsch et al., 2025)	96.4	96.8	88.6	60.2	85.5
$\pi_{0.5}$ (Physical Intelligence et al., 2025)	**98.8**	98.2	98.0	92.4	96.9
OpenVLA-OFT (Kim et al., 2025)	97.6	98.4	97.9	94.5	97.1
GR00T N1 (NVIDIA, 2025)	94.4	97.6	93.0	90.6	93.9
LAPA (Ye et al., 2024)	73.8	74.6	58.8	55.4	65.7
UniVLA (Bu et al., 2025)	96.5	96.8	95.6	92.0	95.2
villa-X (Chen et al., 2025a)	97.5	97.0	91.5	74.5	90.1
VLA-JEPA (Sun et al., 2026)	96.2	**99.6**	97.2	95.8	97.2
PHYSLA (ours)	98.2	**99.6**	**98.4**	**96.6**	**98.2**

Table 3: Success rates (%) on five real-world AgiBot G1 manipulation tasks.

Method	Pick	Flower	Disk	Cup	Plug	Avg.
$\pi_{0.5}$ (Physical Intelligence et al., 2025)	80.0	73.3	46.7	40.0	6.7	49.3
VLA-JEPA (Sun et al., 2026)	82.5	76.7	53.3	36.7	0.0	49.8
PHYSLA (ours)	**87.5**	**93.3**	**73.3**	**60.0**	**13.3**	**65.5**

with the largest gains over VLA-JEPA on Cup (+23.3 percentage points) and Disk (+20.0). Each method is evaluated in 30 trials per task, except Pick, which uses four objects with ten trials each and randomized positions. Figure 4 shows the five tasks from the first-person view. Task definitions are provided in Appendix D.

4.6 ABLATION STUDY

We route physical loss to robot videos (Robot-only), human videos (Human-only), both (Full), or neither (w/o physical), retaining both pretraining streams and the same frozen target generators (Table 10). Full reaches 98.2% average LIBERO success, compared with 96.7% without physical supervision, 97.5% with robot-only supervision, and 97.2% with human-only supervision (Table 4). Both single-source conditions improve over w/o physical. Full exceeds Robot-only by 0.7 percentage points and Human-only by 1.0 point, attaining the highest score on three suites and tying Robot-only on Object. These results favor using physical supervision from both video streams under the evaluated training procedure.

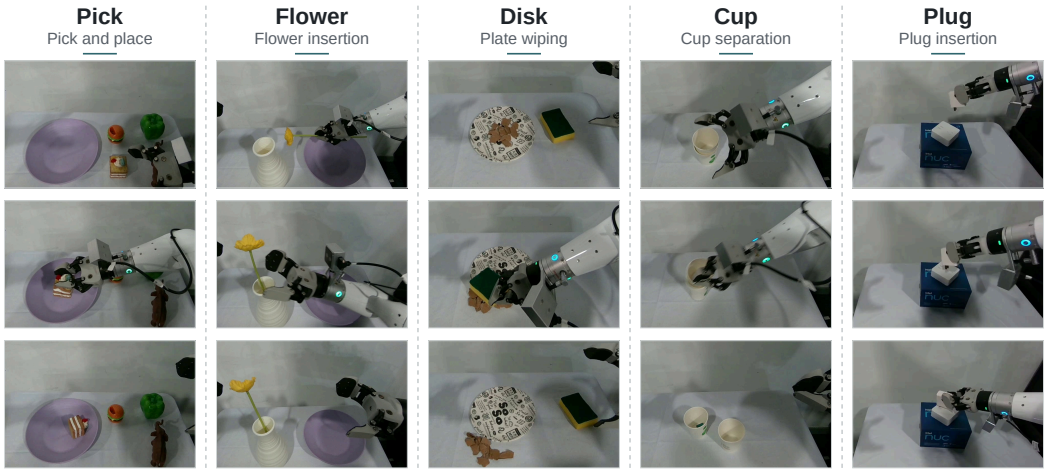

Figure 4: **AgiBot G1 manipulation sequences during test.** First-person frames progress from top to bottom within each task.

Table 4: LIBERO source ablations (%), retaining both pretraining streams. Avg.: mean over four suites. PHYSLA corresponds to the Full variant. Appendix C.

Condition	Spatial	Object	Goal	Long	Avg.
w/o physical	96.2	99.4	96.6	94.6	96.7
Robot-only	97.4	**99.6**	97.6	95.4	97.5
Human-only	97.0	99.4	97.2	95.2	97.2
Full	**98.2**	**99.6**	**98.4**	**96.6**	**98.2**

5 Conclusion

We show that physical supervision can extend beyond the instrumented interactions from which it is acquired. **PhysLA** learns horizon-structured targets from limited physical measurements, infers them from causal RGB, and uses the inferred targets alongside future visual prediction to supervise latent actions across human and robot videos. Representation analyses characterize the temporal and physical information retained by these targets, while source ablations and downstream evaluations support their utility for manipulation. More broadly, our findings position physical measurements as a reusable source of supervision for video-based policy learning without requiring physical sensing at deployment. Expanding both the paired corpus and the collection of uninstrumented videos offers a promising direction for broadening the physical interactions covered by this framework.

Limitations. Our evaluation examines the decodability of selected physical attributes and the utility of physical supervision for downstream control. However, it does not systematically characterize how pressure, tactile deformation, and joint torque encode shared or distinct aspects of interaction within the latent space, or how these relationships are preserved in policy latent actions. Understanding this internal organization under more complex contact dynamics remains an open direction.

AI Use Statement

We used generative AI tools to assist with language editing, polishing, and drafting portions of the manuscript based on technical content, experimental results, and ideas provided by the authors. These tools were not used to conduct experiments, generate datasets, or establish mathematical claims. All AI-assisted text was critically reviewed, revised, and verified by the authors, who take full responsibility for the technical accuracy, citations, claims, and final content of the paper.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Jianxin Bi, Kevin Yuchen Ma, Ce Hao, Mike Zheng Shou, and Harold Soh. VLA-Touch: Enhancing vision-language-action model with dual-level tactile feedback. *IEEE Robotics and Automation Letters*, 11(7):8487–8494, 2026. doi: 10.1109/LRA.2026.3692345.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Learning to act anywhere with task-centric latent actions. In *Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.014.

Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H. Adelson, and Sergey Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4), 2018. doi: 10.1109/LRA.2018.2852779.

Xiaoyu Chen, Hangxing Wei, Pushi Zhang, Chuheng Zhang, Kaixin Wang, Yanjiang Guo, Rushuai Yang, Yucen Wang, Xinquan Xiao, Li Zhao, Jianyu Chen, and Jiang Bian. villa-X: Enhancing latent action modeling in vision-language-action models. *arXiv preprint arXiv:2507.23682*, 2025a.

Yi Chen, Yuying Ge, Weiliang Tang, Yizhuo Li, Yixiao Ge, Mingyu Ding, Ying Shan, and Xihui Liu. Moto: Latent motion token as the bridging language for learning robot manipulation from videos. In *IEEE/CVF International Conference on Computer Vision*, 2025b. doi: 10.1109/ICCV51701.2025.01837.

Zichen Jeff Cui, Hengkai Pan, Aadhithya Iyer, Siddhant Haldar, and Lerrel Pinto. DynaMo: In-domain dynamics pretraining for visuo-motor control. *arXiv preprint arXiv:2409.12192*, 2024. URL https://arxiv.org/abs/2409.12192.

Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 653–660. IEEE, 2024.

Ruoxuan Feng, Jiangyu Hu, Wenke Xia, Tianci Gao, Ao Shen, Yuhao Sun, Bin Fang, and Di Hu. AnyTouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. *arXiv preprint arXiv:2502.12191*, 2025.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *IEEE International Conference on Computer Vision*, 2017.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.16828.

Carolina Higuera, Akash Sharma, Chaithanya Krishna Bodduluri, Taosha Fan, Patrick Lancaster, Mrinal Kalakrishnan, Michael Kaess, Byron Boots, Mike Lambeta, Tingfan Wu, and Mustafa Mukadam. Sparsh: Self-supervised touch representations for vision-based tactile sensing. *arXiv preprint arXiv:2410.24090*, 2024.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025. URL https://arxiv.org/abs/2505.11709.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Chen, Kirsty Ellis, et al. DROID: A large-scale in-the-wild robot manipulation dataset. In *Robotics: Science and Systems*, 2024.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

Junyi Ma, Erhang Zhang, Haoran Yang, Ditao Li, Chenyang Xu, Guangming Wang, and Hesheng Wang. Robot learning from human videos: A survey. *arXiv preprint arXiv:2604.27621*, 2026.

Shangchen Miao, Ningya Feng, Jialong Wu, Ye Lin, Xu He, Dong Li, and Mingsheng Long. JEPA-VLA: Video predictive embedding is needed for VLA models. *arXiv preprint arXiv:2602.11832*, 2026. URL https://arxiv.org/abs/2602.11832.

NVIDIA. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Open X-Embodiment Collaboration. Open X-embodiment: Robotic learning datasets and RT-X models. In *IEEE International Conference on Robotics and Automation*, 2024. doi: 10.1109/ICRA57147.2024.10611477.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. FAST: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Ryan Punamiya, Simar Kareer, Zeyi Liu, Josh Citron, Ri-Zhao Qiu, Xiongyi Cai, Alexey Gavryushin, Jiaqi Chen, Davide Liconti, Lawrence Y. Zhu, et al. EgoVerse: An egocentric human dataset for robot learning from around the world. *arXiv preprint arXiv:2604.07607*, 2026. URL https://arxiv.org/abs/2604.07607.

Sandeep Routray, Hengkai Pan, Unnat Jain, Shikhar Bahl, and Deepak Pathak. ViPRA: Video prediction for robot actions. *arXiv preprint arXiv:2511.07732*, 2025. URL https://arxiv.org/abs/2511.07732.

Vlad Sobal, Wancong Zhang, Kyunghyun Cho, Randall Balestriero, Tim G. J. Rudner, and Yann LeCun. Learning from reward-free offline data: A case for planning with latent dynamics models. *arXiv preprint arXiv:2502.14819*, 2025. URL https://arxiv.org/abs/2502.14819.

Jingwen Sun, Wenyao Zhang, Zekun Qi, Shaojie Ren, Zezhi Liu, Hanxin Zhu, Guangzhong Sun, Xin Jin, and Zhibo Chen. VLA-JEPA: Enhancing vision-language-action model with latent world model. *arXiv preprint arXiv:2602.10098*, 2026.

Kuniyuki Takahashi and Jethro Tan. Deep visuo-tactile learning: Estimation of tactile properties from images. In *IEEE International Conference on Robotics and Automation*, 2019. doi: 10.1109/ICRA.2019.8794285.

Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. In *Advances in Neural Information Processing Systems*, 2022. doi: 10.52202/068431-0587.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Lidén, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*, 2024.

Chi Zhang, Penglin Cai, Ziheng Xi, Haoqi Yuan, Hao Luo, Wanpeng Zhang, Sipeng Zheng, Chaoyi Xu, and Zongqing Lu. Human-centric transferable tactile pre-training for dexterous robotic manipulation. *arXiv preprint arXiv:2607.01067*, 2026a.

Chubin Zhang, Jianan Wang, Zifeng Gao, Yue Su, Tianru Dai, Cai Zhou, Jiwen Lu, and Yansong Tang. CLAP: Contrastive latent action pretraining for learning vision-language-action models from human videos. *arXiv preprint arXiv:2601.04061*, 2026b.

Chuheng Zhang, Tim Pearce, Pushi Zhang, Kaixin Wang, Xiaoyu Chen, Wei Shen, Li Zhao, and Jiang Bian. What do latent action models actually learn? *arXiv preprint arXiv:2506.15691*, 2025.

Wenkang Zhang, Chengbo Yuan, Zicheng Zhang, Zhengxue Cheng, and Yang Gao. EgoTac: In-the-wild tactile prediction from egocentric vision. *arXiv preprint arXiv:2608.15060*, 2026c.

Jialiang Zhao, Yuxiang Ma, Lirui Wang, and Edward H. Adelson. Transferable tactile transformers for representation learning across diverse sensors and tasks. *arXiv preprint arXiv:2406.13640*, 2024.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024. URL https://arxiv.org/abs/2411.04983.

Jianyi Zhou, Ziteng Gao, Feiyang Hong, Zirui Liu, Guannan Zhang, Weisheng Dai, Ruichen Zhen, Chuqiao Lyu, Haotian Wu, Yinian Mao, Xushi Wang, Yuxiang Jiang, Wenbo Ding, and Shuo Yang. TouchAnything: A dataset and framework for bimanual tactile estimation from egocentric video. *arXiv preprint arXiv:2605.13083*, 2026. URL https://arxiv.org/abs/2605.13083.

A Data and Training Details

Recording splits precede window extraction. Source-specific training medians and interquartile scales normalize measurements before differencing. Pressure grids are validity-weighted pooled from 21×21 to 7×7 per hand; marker displacements use a 12×16 coordinate chart; torque uses six joint channels. Table 5 distinguishes physical-encoder training windows from the paired windows retained after RGB decoding checks for supervision transfer. Nine RH20T training windows are excluded because their RGB frames cannot be decoded; validation counts are unchanged.

Table 5: Training and validation windows. RGB train counts exclude decoding failures.

Source	Measurement	Physical train	RGB train	Validation
EgoTouch	Pressure	104,560	104,560	10,174
RH20T/Franka	Joint torque	89,369	89,360	17,218
VLA-touch	Marker displacement	13,706	13,706	1,490

Table 6: Architecture and policy-pretraining configuration.

Component	Configuration
Physical targets	8 samples; 3/3/2 temporal groups; 128 dimensions per group
Physical encoder	13-to-192-dimensional MLP; 2 Transformer layers, 6 heads, FFN width 768; shared horizon projection
Measurement decoder	7-to-192-dimensional query embedding; MLP widths $320 \to 192 \to 2$
RGB predictor	Frozen DINOv2 patch features; width 192, 6 heads, 2 temporal layers, FFN width 768; one spatial query per frame and 3 horizon queries
RGB input	8 frames at or before t
VLA backbone	Qwen3-VL-2B-Instruct (Bai et al., 2025)
Policy tokens	24 latent-action tokens in 3 groups; 32 embodied-action tokens
Visual teacher	V-JEPA 2 ViT-L, `vjepa2-vitl-fpc64-256` (Assran et al., 2025); 8 frames, 4 tubelets
Physical prediction head	Width 256, 2 layers, 8 heads, FFN width 1024; one summary token per horizon
Policy optimizer	AdamW, $(\beta_1, \beta_2) = (0.9, 0.95)$
Policy learning rates	Backbone 10^{-5}; action decoder 10^{-4}; other trainable modules 3×10^{-5}
Policy batch and budget	16 robot and 16 human samples per device; 50,000 iterations

Physical encoding. The encoder is trained for 20,000 updates with global batch size 16, AdamW, learning rate 3×10^{-4}, weight decay 0.05. Inverse-frequency sampling gives the three sources equal expected weight. The masking routine uses 25% for sequences of at most 96 valid tokens and 35% otherwise, selecting contiguous token spans or spatial neighborhoods, respectively. Reconstruction averages masked valid scalar components. The checkpoint minimizes validation reconstruction loss.

RGB-to-physical transfer. With E_{phys} and DINOv2 frozen, we train G_{RGB} for 30,000 updates using AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay 0.05, and global batch size 16 (2 per GPU on 8 GPUs, without gradient accumulation). The learning rate increases linearly to 3×10^{-4} over 2,000 warmup updates and then decays by cosine to 10^{-6}; the gradient norm is clipped at 1.0. We use seed 20260910. Sources are sampled in a 1:1:1 ratio, cycling through a shuffled list of windows within each source. The RGB training set contains 207,626 valid paired windows (Table 5); each input comprises eight causal frames. Training minimizes $\mathcal{L}_{\text{transfer}}$ with target normalization statistics fitted on the training split. Every 5,000 updates, we evaluate a fixed validation subset of 1,024 windows per source, sampled without replacement using the same seed. For each source, validation loss averages windows, horizons, and feature dimensions; the selection score averages the three source losses equally. We select the checkpoint with the lowest score, preferring the earlier checkpoint in a tie. Test data are excluded from model selection.

Temporal sampling. EgoTouch uses every fourth frame at approximately 30 Hz. Relative to the last history frame t, the eight history samples span $[-0.933, 0]$ s and the eight physical samples span $[0.133, 1.067]$ s. Robot windows use every second synchronized row within an episode. Physical sample i has normalized time $i/7$; groups contain indices $\{0, 1, 2\}$, $\{3, 4, 5\}$, and $\{6, 7\}$. Visual

targets comprise four V-JEPA 2 tubelets and three successive transitions. Their ordering associates them with the three physical groups; their time spans need not coincide.

Policy pretraining. Each iteration performs one robot and one human optimizer update, giving 100,000 updates with no gradient accumulation. The scheduler advances per update. The DROID plan contains 128,000 draws from 127,645 distinct windows. After excluding 994 undecodable video windows, the human stream contains 82,747 EgoDex, 1,141 EgoTouch, and 43,118 SSV2 windows, sampled uniformly within the combined manifest (65.15%, 0.90%, and 33.95%). Each retained sample is joined by its unique identity to a finite 3×128 physical target; missing identities cause an error. The retained human manifest has 127,006/127,006 target matches; this coverage is conditional on decoding and subset selection.

B Physical Evaluation

B.1 Representations and Splits

Table 7: Probe inputs and physical supervision.

Representation	Input	Physical supervision
DINOv2	Causal RGB	None
G_{RGB}	Causal RGB	Frozen physical targets
PS-Prehead	Past RGB + hand pose	Pressure-map prediction
PressAlign	Past RGB + hand pose	Pressure latents and decoder consistency
E_{phys}	Physical measurements	Masked reconstruction

DINOv2 concatenates framewise patch means; G_{RGB} learns spatial and temporal aggregation from the same RGB history. PS-Prehead averages TouchAnything's RGB–pose features over joints and frames per horizon (Zhou et al., 2026); PressAlign uses pressure-latent queries. E_{phys} receives unmasked measurements as a privileged reference.

EgoTouch follows its official seen/unseen labels after validity filtering. The paired training and validation sets contain 1,438 and 173 disjoint recordings. Probe test recordings overlap neither set nor the EgoTouch policy-pretraining subset. The nine unseen tasks are absent from these EgoTouch training and validation subsets. The official seen set retains 213 windows from `pick_up_cotton_swab` and `take_billiard_ball`, which have no original training recordings. The pose-valid panel contains 8,278/2,632 seen/unseen windows; policy probes use 12,893/4,075, with 83 unseen recordings. RH20T and VLA-touch use recording-disjoint test splits.

B.2 Probe Protocol

Pressure labels use valid 21×21 measurements clipped to $[0, 1]$. Let m_i average pressure equally across hands and c_i be the corresponding mean fraction of cells exceeding 0.1. Continuous labels average m_i, $d_i = m_i - m_{i-1}$, and $|d_i|$ per horizon. Contact is $b_i = \mathbf{1}[c_i > 0.01]$; onset/release are $b_i(1 - b_{i-1})/(1 - b_i)\, b_{i-1}$, maximized per horizon. Marker labels average d_x, d_y, and $\|\Delta d\|$ over valid markers and time; torque labels are per-joint level, signed change, and absolute change. Differences require both readings to be valid. Initial differences and transition events are zero. Changes precede pooling, assigning boundary events to the later horizon.

Independent ridge probes fit each representation/source/horizon/attribute using training-standardized features, summed squared error, and an unpenalized intercept; constant features use unit scale. Validation R^2 or AUROC selects $\alpha \in \{0.01, 0.1, 1, 10, 100, 1000\}$ for continuous or event labels, respectively; ties favor larger α. Probes are not refit on validation data, and the same fits evaluate seen and unseen tests. Scores average three horizons and, for torque, six joints.

Qualitative decoding. Figure 3 averages both hands over the first horizon's three samples. Active area uses a 0.03 pressure threshold, distinct from the contact label. Points use mean measurement timestamps; lines connect discrete estimates without smoothing.

Policy probes. **PHYSLA** and w/o physical use the controlled Full and w/o physical checkpoints from Appendix C, with matched initialization, data sampling, optimizer settings, and update budgets. Both receive identical current images and task instructions. We extract the final backbone hidden states at the 24 latent-action slots. For each horizon, its eight 2,048-dimensional tokens are concatenated in slot order into a 16,384-dimensional feature vector. Independent ridge probes fit each horizon and attribute using the standardization and validation-based regularization protocol above. The backbone remains frozen, and the trained physical prediction head is not used to construct probe features.

B.3 TEMPORAL RECONSTRUCTION

We evaluate the selected 20,000-update physical encoder and its decoder on all 28,882 validation windows: 10,174 EgoTouch, 17,218 RH20T, and 1,490 VLA-touch windows. Evaluation uses the training masking rule, with one mask sampled per window and shared across all four conditions. The masking generator uses seed 20261911; windows are processed in source order (EgoTouch, RH20T, VLA-touch), preserving manifest order within each source, with batch size 16. The reported values use one masking pass, without averaging over repeated mask draws.

Reflection replaces τ by $1 - \tau$ in encoder inputs and decoder queries while retaining measurement values and token order. Permutation supplies $(y^{*,2}, y^{*,3}, y^{*,1})$ to unchanged decoder queries; the zero reference predicts zero in normalized measurement space. For each condition, SmoothL1 with transition parameter 1 is summed over all masked, valid scalar measurement components and divided by their total count across windows, horizons, and sources. Thus each valid masked scalar has equal weight; the three sources are not averaged equally. The resulting ordered/zero/reflected/permuted losses are 0.3323/0.7050/0.4038/1.2317, respectively.

B.4 ADDITIONAL PROBE RESULTS

Table 8: Additional EgoTouch representations on the panel in Table 1(a): R^2 for continuous attributes and AUROC for events.

Representation	Pressure	Change	|Change|	Contact	Onset	Release
(a) Seen: 8,278 evaluation windows						
PressAlign	0.6238	0.2954	0.2913	0.9445	0.8237	0.8136
E_{phys} (priv.)	**0.9961**	**0.9707**	**0.6951**	**0.9914**	**0.9609**	**0.9618**
(b) Unseen: 2,632 evaluation windows						
PressAlign	0.4948	0.0539	0.1288	0.9355	0.8477	0.8548
E_{phys} (priv.)	**0.9960**	**0.9718**	**0.5972**	**0.9940**	**0.9680**	**0.9754**

Table 9: Robot physical probes (R^2).

Attribute	G_{RGB}	DINOv2	E_{phys} (priv.)
(a) RH20T torque: 20,623 evaluation windows			
Mean torque	0.3264	-0.1393	0.9826
Signed change	0.2852	-0.0588	0.8538
Absolute change	0.1321	-0.1583	0.7344
(b) VLA-touch marker displacement: 1,676 evaluation windows			
d_x	0.4286	-0.1282	0.9875
d_y	0.4918	-0.0519	0.9906
$\|\|\Delta d\|\|$	0.2747	-0.0349	0.9685

C SOURCE ABLATIONS

Table 10 changes only physical-loss routing across pretraining and adaptation. Initialization, data, world/action objectives, budgets, and the enabled physical-loss configuration are shared. These comparisons assess the full training procedure.

Table 10: Physical-supervision routing. Both pretraining data streams are retained.

Condition	Human pretraining	Robot pretraining	LIBERO adaptation
w/o physical	Off	Off	Off
Robot-only	Off	On	On
Human-only	On	Off	Off
Full	On	On	On

D Downstream Evaluation

LIBERO. We sample the four no-op-filtered LIBERO suites equally, excluding LIBERO-90. Full uses action/world/physical losses for 30,000 updates, with R_{phys} frozen and gradients propagated through it. The backbone, world predictor, and action decoder update at learning rates 10^{-5}, 3×10^{-5}, and 10^{-4}, respectively, with 5,000 warmup updates and cosine decay to 10^{-6} The final checkpoint is evaluated on 50 episodes per task with common initial states and official success criteria.

SimplerEnv. Adaptation uses Bridge and RT-1 (`bridge_orig`, `fractal20220817_data`) equally for 30,000 updates, without simulation demonstrations. Visual matching uses embodiment-specific action normalization and controllers. Each embodiment's average weights its four task groups equally.

G1. Methods share demonstrations, global batch size 256, and 20,000 adaptation updates, and use final checkpoints. VLA-JEPA starts from its released weights. Each method has 40 Pick trials (10 per object, randomized positions) and 30 per remaining task; the reported average weights five tasks equally.

Reporting. Success rates summarize rollouts of the evaluated checkpoints. They do not estimate variability across independent pretraining or adaptation runs. Probe intervals, when reported, concern fixed-checkpoint evaluation samples and are distinct from training-seed variation.

Table 11: G1 success criteria and evaluation variations.

Task	Success criterion	Variation
Pick	Lift and release the object stably in the target region	Object, position
Flower	Insert the stem; the holder supports it after release	Holder pose
Disk	Wipe the designated region with continuous contact; keep the plate in place	Plate pose, wipe direction
Cup	Release one separated cup stably in the target region; leave the stack standing	Stack, target positions
Plug	Insert to the designated depth without dislodging the socket	Socket pose

D.1 Additional Simulation Results

Table 12: SimplerEnv success rates (%) under visual matching. †: fine-tuning with simulation demonstrations.

Google Robot					
Method	Pick	Move	Drawer	Place	Avg.
π_0 (Black et al., 2024)	72.7	65.3	38.3	–	–
π_0-fast (Pertsch et al., 2025)	75.3	67.5	42.9	–	–
GR00T N1 (NVIDIA, 2025)	0.7	1.9	2.9	0.0	1.4
villa-X (Chen et al., 2025a)	81.7	55.4	38.4	4.2	44.9
Moto (Chen et al., 2025b)	74.0	60.4	43.1	–	–
VLA-JEPA (Sun et al., 2026)	88.3	64.1	59.3	49.1	65.2
Ours (Full)	**88.8**	**76.2**	**68.3**	**62.7**	**74.0**
WidowX					
Method	Spoon	Carrot	Block	Eggplant	Avg.
π_0 (Black et al., 2024)	29.1	0.0	16.6	62.5	–
π_0-fast (Pertsch et al., 2025)	29.1	21.9	10.8	66.7	–
OpenVLA-OFT (Kim et al., 2025)	34.2	30.0	30.0	72.5	41.8
GR00T N1 (NVIDIA, 2025)	1.4	0.0	0.0	13.9	3.8
LAPA† (Ye et al., 2024)	70.8	45.8	54.2	58.3	57.3
UniVLA (Bu et al., 2025)	–	–	–	–	42.7
villa-X (Chen et al., 2025a)	48.3	24.2	19.2	71.7	40.8
VLA-JEPA (Sun et al., 2026)	75.0	70.8	12.5	70.8	57.3
Ours (Full)	**81.2**	**75.3**	**23.5**	**78.6**	**64.7**