HapticWAM: Distilling Imagined Touch into a World–Action Model without Inference-Time Tactile Sensing

Mikhail Sannikov*, Ilya Mikhalchuk*, Konstantin Gubernatorov*, Petr Kovalev,
Ogunwoye Faith Oluwatobi, and Dzmitry Tsetserukou

Project page: https://advanced-robotic-manipulation.github.io/websites/hapticwam/

***Abstract*— Contact-rich manipulation requires estimating forces, slip and contact geometry that can remain ambiguous in scene images. Optical tactile sensors provide both visual observations of the contact surface and mechanical measurements, yet learning from these signals raises two challenges: representing contact beyond appearance and transferring its benefits to a policy that does not require fingertip observations at deployment. We introduce HapticWAM, a world–action model that combines heterogeneous tactile encoding, structured contact prediction and teacher–student distillation. Its teacher encodes gel images together with deformation, shear, distributed forces, resultant wrench and derived contact state into a frozen video backbone. Rather than predicting tactile pixels alone, the model jointly generates actions and a contact package describing future events and mechanics. Anticipatory Contact Coupling uses the previously imagined package to condition attention, preserving a contact-related input when direct tactile observations are unavailable. Haptic-Imagination Distillation transfers both contact futures and action predictions to a student that retains the generative contact head but removes its fingertip input branches. On a real-world setup, across three contact-rich pick-and-place tasks, HapticWAM Student achieves a 77% per-task mean success rate (41 of 50 starts, 82% pooled), reaching 95% on one of the tasks, outperforming the evaluated teacher and baseline configurations.**

I. Introduction

Picking up a fragile object requires more than reaching the correct pose. The fingers must establish contact, maintain enough force to prevent slip and release the object at the target. Scene images reveal object geometry and motion but may not distinguish a stable grasp from light contact, particularly when the gripper occludes the interaction. Optical tactile sensing complements vision with local contact appearance and mechanics. The central question is how to learn from this richer information while reducing the deployed policy's dependence on direct fingertip measurements.

Vision–language–action models (VLAs) [1]–[4] map observations and task instructions to action chunks. Tactile VLAs add contact observations to this interface [5]–[9]. These policies can learn anticipatory behavior without an explicit world model. However, action supervision alone does not require them to expose a prediction of the contact events or mechanical changes associated with a candidate motion. World–action models (WAMs) jointly generate actions and future observations [10]–[14], providing a framework in which these quantities can be learned explicitly. Predicting visual futures by itself does not directly supervise grip force, contact geometry or slip.

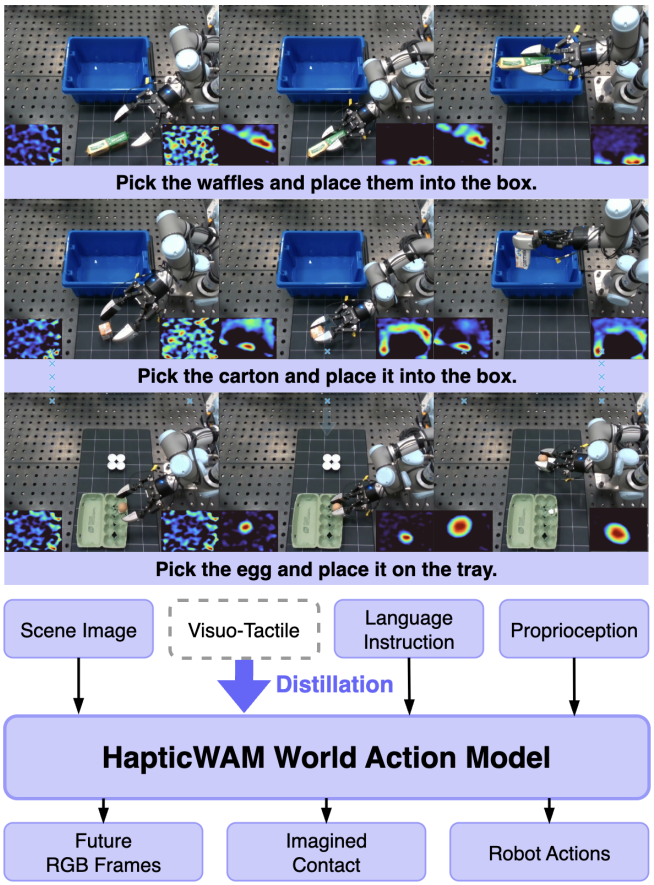

Fig. 1. HapticWAM encodes the full range of contact modalities an optical-tactile sensor reports, rather than gel appearance alone. Our approach leverages Haptic Imagined Distillation to preserve inference-time tactile-aware reasoning while eliminating optical-tactile sensors as input through distillation.

For optical tactile sensors, the available information also extends beyond an image of the gel. Native outputs include deformation, indentation, shear, distributed forces and resultant wrench; contact masks, centres of pressure and slip indicators can be derived from these streams. An appearance-only input leaves this mechanical structure implicit in the image representation. Conversely, reconstructing future tactile pixels allocates prediction capacity to appearance as well as task-relevant contact changes. Prior tactile world models

*Denotes equal contribution.

All authors are with the Intelligent Space Robotics Laboratory, Skolkovo Institute of Science and Technology, Moscow, Russia. {Mikhail.Sannikov, Ilia.Mikhalchuk, Konstantin.Gubernatorov, Petr.Kovalev, Faith.Ogunwoye,D.Tsetserukou} @skoltech.ru

already explore image-based and structured predictions [15]–[18]. We build on this progression by connecting heterogeneous native sensor inputs to an explicit contact-future representation that survives removal of tactile observations from the learned policy.

This removal introduces a second problem. A policy trained to condition on measured contact cannot simply retain the same observation interface when those measurements are unavailable. Privileged-modality distillation offers a way to transfer supervision from a tactile teacher to a student with fewer inputs. Here the student must predict contact from visual approach geometry, robot state and an arm-side wrench estimate. It cannot recover arbitrary unobserved contact conditions, but it can be trained to represent the teacher's predicted contact consequences alongside its actions. Matching these futures provides an explicit supervision target beyond copying actions; whether it improves training over action-only distillation is a separate empirical question.

We propose **HapticWAM** to connect these representation and transfer problems. A Heterogeneous Haptic Tokenizer (HHT) maps tactile appearance and mechanics into a frozen video diffusion transformer. The model jointly generates an action chunk and a structured package of contact events, mechanical changes and contact geometry. Anticipatory Contact Coupling (ACC) conditions attention on the package imagined at the previous replan, so its contact-related input need not come from a live tactile stream. Haptic-Imagination Distillation (HID) then matches the teacher's contact futures and action predictions in a student that removes its tactile observation frames but retains contact generation. Figure 1 illustrates the evaluated tasks. The deployment claim concerns tactile *policy inputs*: the robot comparisons retain mounted sensors and the shared controller's tactile guards.

Our contributions are:

1) **Heterogeneous Haptic Tokenization** combines gel appearance, mechanical fields and derived contact state (the full modality range available from DM-Tac W2L sensors) in a common latent-frame interface, supporting joint structured-contact and action prediction.
2) **Anticipatory Contact Coupling** uses current leading signals and the previous imagined contact package to bias attention toward haptic tokens, including when fingertip observations are unavailable.
3) **Haptic-Imagination Distillation** combines teacher initialization, future-contact and event matching, action matching and ground-truth supervision in a student without tactile inputs.
4) **A visuo-tactile dataset and real-wrold robot evaluation** support comparisons of the teacher, student and visual baselines on shared pick-and-place starts, together with a deployment-time contact-generation intervention.

II. Related Work

A. World–Action Models

Imitation policies [19] and VLAs [2], [3] predict actions from observations and task context. World–action models add an explicit generative future. Video foundation models [10] support zero-shot robot control [11], value-guided generation [12] and joint predictive reasoning and action learning [13]. Efficient inference is an active direction [14], and Being-H0.7 goes further by replacing frame generation with learnable latent queries between perception and action, arguing that pixel-space prediction is a costly and indirect substrate for control [20]. Whether the future is rendered as pixels or compressed into latents, it is learned from video: a visually plausible future does not by itself specify the forces and contact transitions needed to execute it. This distinction matters when different contact states produce similar scene images.

HapticWAM shares the view that pixel fidelity is not the objective, but rather than removing the explicit future it changes what that future describes: the pretrained visual backbone is retained while the prediction space is extended with contact events, force and displacement changes, and contact geometry. Video provides scene context, whereas contact streams provide direct targets for the interaction quantities represented by the model. The distinction is a choice of supervision and representation, not a claim that action-only policies cannot learn anticipatory behavior or that explicit contact generation alone guarantees better control.

B. Tactile World Models

Dream-Tac combines tactile prediction with contact-aware attention [15], while Tactile-WAM uses asymmetric attention over tactile tokens [16]. VT-WAM and TacForeSight couple visual–tactile prediction with force guidance [21], [22]; TacPAC corrects actions using discrepancies between expected and observed touch [23]. Other approaches predict tactile futures for planning [24], [25], imagine touch in a VLA [26], or model wrist forces [27]. Structured prediction is also established: N_0-TWAM uses a force-based representation [17], and HiTac-WAM factorizes contact, deformation and slip [18]. These precedents motivate treating touch as more than a stream of images.

Our focus is the connection between heterogeneous tactile measurements, future-contact generation and deployment without direct tactile policy inputs. HHT gives appearance, mechanical fields and contact state distinct encoding paths rather than requiring a single image encoder to infer all of them. The output package represents contact events and mechanical changes instead of gel appearance alone. ACC then uses the model's previously predicted package, making this attention-conditioning input available to the student after its fingertip observation branches are removed. The proposed contribution is this combination, not the first use of force prediction, structured touch or contact-aware attention.

C. Privileged-Modality Distillation

Privileged teachers transfer information unavailable to deployed students. In manipulation, PTLD transfers tactile representations for sim-to-real control [28], FD-VLA distills force supervision [29], and TacImag predicts tactile observations from vision and proprioception [30]. HapticVLA [9]

distills a tactile-conditioned reactive policy into a student that predicts a tactile token. These methods address observation asymmetry, but differ in what the student retains: a teacher representation, an imitated action or a predicted sensory quantity.

HID retains a generative *contact-future* head and supervises its mechanics and event predictions alongside the action field. Event saliency and predicted teacher uncertainty weight contact matching, while ground-truth supervision anchors the student to demonstrations. Recorded teacher and earlier-student rollouts broaden the training distribution, following the dataset-aggregation motivation [31]. Fully on-policy distillation, in which the teacher supervises samples the student itself generates, outperforms off-policy matching in language-model settings and, with reward extrapolation, lets a student exceed its teacher [32]; HID approximates this setting offline from recorded rollouts rather than collecting new on-policy data for each student, and does not extrapolate beyond the teacher. This gives the student an explicit predicted-contact interface without requiring fingertip observations. The present robot ablation tests use of that interface at deployment, rather than isolating the training benefit of contact targets over action-only matching.

III. Method

A. Problem Formulation

At each replan t, the policy receives a scene image I_t, robot state q_t, task text and an arm-side wrench history $w_{t-k:t}$ over the last k control ticks. The wrist signal is estimated from motor currents, not measured by a wrist force sensor. The teacher additionally observes gel images G_t^f, mechanical fields X_t^f and derived contact states κ_t^f from two fingertips indexed by $f \in \{1,2\}$; superscripts T and S denote teacher and student throughout:

$$o_t^{\mathrm{T}} = (I_t, q_t, w_{t-k:t}, \{G_t^f, X_t^f, \kappa_t^f\}_f), \tag{1}$$

$$o_t^{\mathrm{S}} = (I_t, q_t, w_{t-k:t}). \tag{2}$$

Task text is supplied as shared conditioning in addition to these sensor inputs. Both policies predict a chunk of end-effector pose deltas and absolute gripper apertures. The student removes tactile observations, while retaining vision, proprioception and the arm-side wrench estimate.

A frozen Cosmos video diffusion transformer [10] processes observation, video, contact and action groups in one latent sequence. Its backbone is a diffusion transformer (DiT) [33]: a stack of 28 identical blocks that denoise spatio-temporal latent tokens with self-attention, cross-attention to the task-text embedding and a feed-forward multilayer perceptron (MLP). The sequence is

$$z = [z_{\mathrm{obs}}, z_{\mathrm{video}}, z_{\mathrm{contact}}, z_{\mathrm{action}}]. \tag{3}$$

Observation frames are fixed conditioning. Generated groups are jointly denoised using rectified flow: a clean latent x_0 is mixed with Gaussian noise ε at noise level $\tau \in [0,1]$, the network v_θ regresses the velocity target $v^\star$, and $\hat{x}_0$ is the denoised estimate:

$$x_\tau = (1-\tau)x_0 + \tau\varepsilon, \tag{4}$$

$$v^\star = \varepsilon - x_0, \tag{5}$$

$$\hat{x}_0 = x_\tau - \tau v_\theta(x_\tau, \tau, o_t). \tag{6}$$

Only low-rank adaptation (LoRA) adapters [34], observation encoders, ACC and the contact readout heads are trainable. The previously executed action chunk also conditions the backbone. Figure 2 summarizes the architecture.

B. Heterogeneous Haptic Tokenization

HHT maps heterogeneous observations into the backbone's latent-frame interface. Each fingertip supplies gel appearance and eight mechanical channels: two deformation components, depth, two shear components and three distributed-force components. Its 11-dimensional contact state contains six resultant-wrench components, area, two centre-of-pressure coordinates, slip and mask fraction. The frozen variational autoencoder (VAE) encodes the gel images; a learned pointwise convolution fuses the two fingertips into one appearance frame. A convolutional encoder $\mathscr{E}_X$ processes mechanical fields with contact-state FiLM modulation [35], whose scale γ and shift β an MLP computes from κ_t^f:

$$z_t^f = (1+\gamma(\kappa_t^f)) \odot \mathscr{E}_X(X_t^f) + \beta(\kappa_t^f), \tag{7}$$

where $\odot$ is the elementwise product. The modulation MLP is zero-initialized. A causal temporal convolution [36] $\mathscr{E}_w$ encodes the wrench history, and an MLP encodes robot state. The student deletes the fused gel and mechanics observation frames, reducing the sequence from 14 to 12 latent frames.

C. Structured Contact Prediction

For each of the $J = 3$ future latent steps $j = 1, \ldots, J$, the contact package contains a typed event e; per-finger changes in the displacement field ΔD^f and normal-force field $\Delta F^{z,f}$ (superscript z: the normal component); a contact mask M^f, centre of pressure ξ^f, slip score s^f and resultant wrench F^f; and the arm-side wrench w (time indices are omitted inside the tuple):

$$c_{t+j} = (e_{t+j}, \{\Delta D^f, \Delta F^{z,f}, M^f, \xi^f, s^f, F^f\}_f, w_{t+j}). \tag{8}$$

Events $e \in \{\texttt{none}, \texttt{onset}, \texttt{hold}, \texttt{slip}, \texttt{release}\}$ distinguish no contact, onset, hold, slip and release. Targets are derived from thresholded recorded fields on the latent-time grid, without manual contact annotation. Predicting changes emphasizes contact transients rather than reconstructing unchanged tactile appearance.

A deterministic codec packs fields and masks on the latent grid, events as one-hot bands, centres of pressure as Gaussian bumps, slip as tiled values and wrenches as spatial blocks. Undefined centres of pressure produce no bump. Actions are

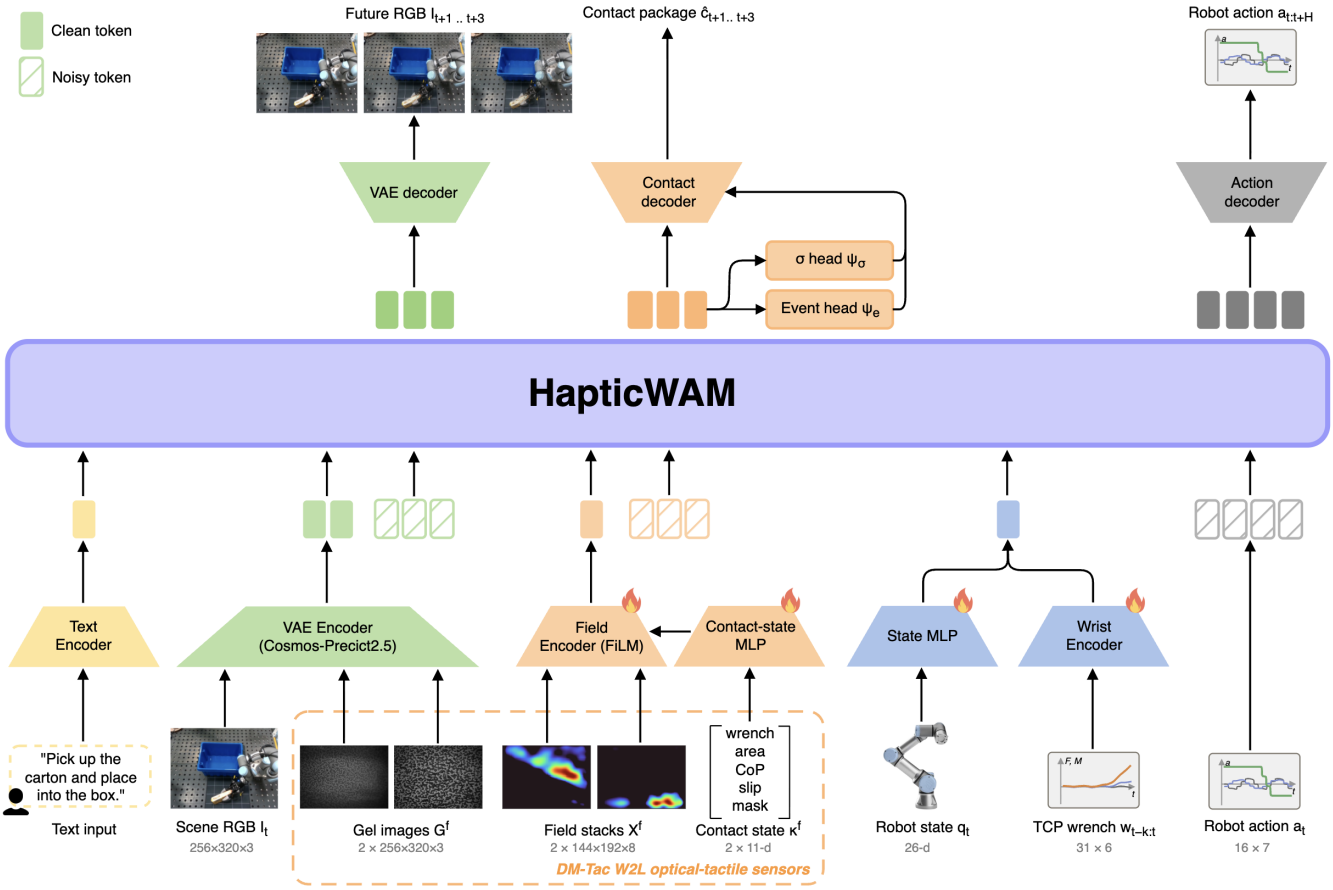

Fig. 2. HapticWAM framework overview. *Inputs:* the scene image I_t, robot state q_t and arm-side wrench history $w_{t-k:t}$, with task text and the previously executed action chunk as conditioning; the teacher additionally observes both fingertips' gel images G_t^f, mechanical fields X_t^f and contact states κ_t^f. Images pass through the frozen video VAE, mechanical fields through a convolutional encoder, and robot state and wrench history through MLP and causal-convolution encoders. *Outputs:* future video, a structured contact package per future step, and a chunk of end-effector pose deltas and gripper apertures. The student is the same model without visuo-tactile input modalities from optical-tactile sensors.

packed as four time steps per frame. Event and uncertainty heads ψ_e and ψ_σ read the contact-frame hidden states h^c:

$$\hat{p} = \mathrm{softmax}(\psi_e(h^c)), \tag{9}$$

$$\log \sigma = \psi_\sigma(h^c). \tag{10}$$

Events and mechanical channels are generated jointly; the event head is not a separate sampling stage. A structural mask blocks direct attention from contact/action queries to generated-video keys. Shared weights and unmasked observation rows still permit indirect coupling.

Contact reconstruction uses a heteroscedastic loss over the output groups m (displacement, normal force, mask, centre of pressure, slip, fingertip wrench and arm-side wrench), each with its own predicted uncertainty σ_m:

$$d_m = \|\hat{x}_0^{(m)} - x_0^{(m)}\|_2^2, \tag{11}$$

$$\mathcal{L}_{\mathrm{con}} = \sum_m (d_m \sigma_m^{-2} + \log \sigma_m^2)\, \mathrm{sg}[\sigma_m^2]^\nu. \tag{12}$$

Here sg stops gradients and the exponent $\nu \in [0,1]$ moderates inverse uncertainty weighting. The teacher objective, with scalar loss weights $\lambda_{(\cdot)}$ given in (26)–(27), is

$$\begin{aligned}\mathcal{L}^{\mathrm{T}} = {} & \lambda_a \mathcal{L}_{\mathrm{act}} + \lambda_v \mathcal{L}_{\mathrm{vid}} + \lambda_c \mathcal{L}_{\mathrm{con}} + \lambda_w \mathcal{L}_w \\ & + \lambda_e(\mathcal{L}_{\mathrm{evt}} + \mathcal{L}_{\mathrm{band}}) + \mathcal{L}_{\mathrm{ACC}} + \lambda_\sigma \|\log \sigma\|_2^2.\end{aligned} \tag{13}$$

Action and video terms match flow velocities; event terms supervise the readout and packed event band; the wrench term reconstructs the arm-side signal. Failed demonstrations have zero ground-truth action weight.

D. Anticipatory Contact Coupling

ACC combines current leading signals with the contact package imagined at the previous replan. A fusion MLP ϕ produces the representation

$$\mathbf{u}_t = \phi\left(\begin{bmatrix} \mathcal{E}_w(w_{t-k:t}) \\ \bar{z}_t^{\mathrm{vis}} \\ \mathrm{MLP}(a^{\mathrm{prev}}) \\ \mathrm{MLP}(\bar{c}_{t+1|t-1}) \end{bmatrix}\right). \tag{14}$$

The visual summary pools conditioning tokens; the teacher additionally uses tactile tokens. The contact summary is aligned to the next replan from the previously generated package, rather than taken from the ground truth. Learned projections W_g and W_e give the anticipatory gate and event distribution

$$g_t^{\mathrm{ant}} = \mathrm{sigmoid}(W_g \mathbf{u}_t), \tag{15}$$

$$p_t^{\mathrm{evt}} = \mathrm{softmax}(W_e \mathbf{u}_t). \tag{16}$$

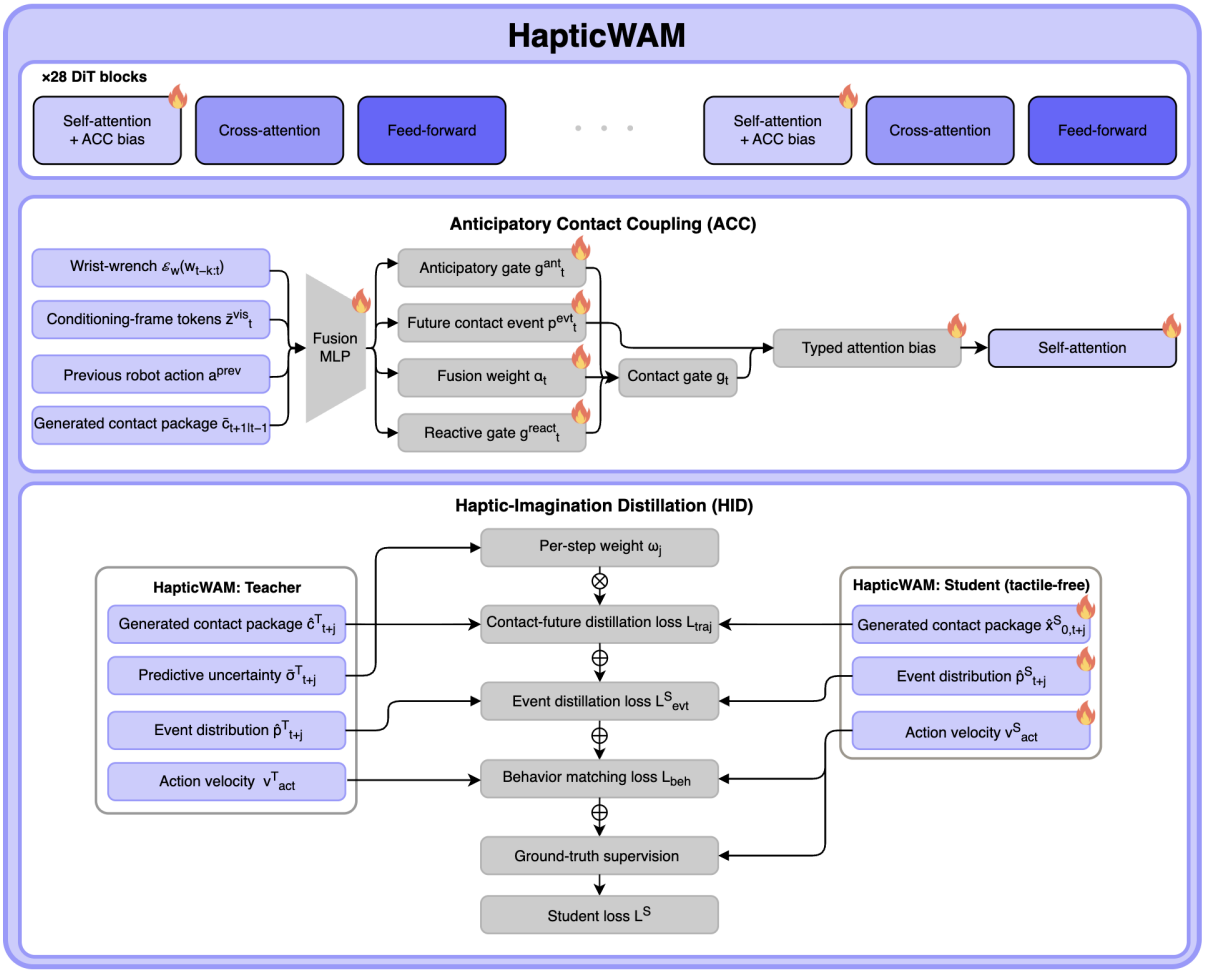

Fig. 3. HapticWAM internals. *Backbone:* the 28 DiT blocks of the frozen video model, each with self-attention, cross-attention to the task-text embedding and a feed-forward MLP. *ACC* fuses the wrench encoding, a pooled visual summary, the previous action chunk and the contact package imagined at the previous replan into an anticipatory gate g_t^{ant}, an event distribution p_t^{evt} and a blending weight α_t; the teacher blends g_t^{ant} with a reactive gate from observed tactile change. The student uses the anticipatory branch alone. Each block adds the typed attention bias, to self-attention scores on haptic keys only. *HID* samples the frozen teacher for contact packages, event distributions, uncertainty and action velocities, and trains the student by contact-future matching weighted by event saliency and teacher uncertainty, event distillation and action-velocity matching.

The teacher blends this gate with a reactive gate derived from observed tactile change:

$$g_t = \alpha_t g_t^{\text{ant}} + (1-\alpha_t) g_t^{\text{react}}. \tag{17}$$

The student uses only the anticipatory branch. In each DiT block ℓ, the pre-softmax attention score $r_{i,n}^{(\ell)}$ between query i and haptic key n receives a typed bias:

$$r_{i,n}^{(\ell)} \leftarrow r_{i,n}^{(\ell)} + \mu_\ell \, b(p_t^{\text{evt}}) \, g_t, \tag{18}$$

$$b(p) = \sum_e p_e \, \text{softplus}(\zeta_e), \tag{19}$$

with learned per-event parameters ζ_e and per-block scales μ_ℓ that start at zero. Haptic keys comprise tactile/contact frames for the teacher and contact/proprioceptive frames for the student. Contact and event labels supervise ACC through binary cross-entropy (BCE) and cross-entropy (CE):

$$\mathscr{L}_{\text{ACC}} = \lambda_g \, \text{BCE}(g_t, y_t) + \lambda_e \, \text{CE}(p_t^{\text{evt}}, e_{t+1}), \tag{20}$$

where $y_t \in \{0,1\}$ indicates contact in the labeled future interval. The teacher's reactive gate g_t^{react} is a learned sigmoid MLP of mean absolute tactile-field change; its blending coefficient α_t is a sigmoid projection of $\mathbf{u}_t$. In two-pass training, a gradient-free sample supplies the previous-package input for both policies; the inner pass uses a noisy target summary to terminate recursion. Deployment reuses the previous prediction, or a self-generated package at the first replan, without future measurements. At deployment, ACC changes attention internally; a separate scripted contact-probability veto controls gripper closure. No physical anticipation lead time is inferred from this architectural design. Figure 3 summarizes the architecture of ACC and HID introduced further in Sec. III-E.

E. Haptic-Imagination Distillation

The student inherits teacher weights but omits tactile observation frames. For each training batch, the frozen teacher generates contact targets, event distributions and predicted uncertainty. A per-step weight ω_j combines event saliency, with coefficient η, and the teacher uncertainty $\bar{\sigma}^{\text{T}}_{t+j}$ (the mean

Algorithm 1 Haptic-Imagination Distillation

Require: Frozen tactile teacher; demonstrations and mixed rollouts

Initialize the student from teacher weights
Remove student gel and mechanics observation frames
for each training minibatch **do**
    Construct paired teacher and student observations
    Sample teacher contact, events and uncertainty without gradients
    Draw shared action noise and a shared noise level
    Predict student contact and action velocities
    Match contact targets using (21)–(23)
    Match teacher action velocities using (24)
    Mask ground-truth action loss on unsuccessful episodes
    Add ground-truth terms and ACC auxiliaries in (25)
    Update student adapters, encoders and heads; update weight EMA
end for
return Student EMA weights for deployment

of σ_m over output groups) with temperature ρ:

$$\omega_j = \big(1+\eta(1-\hat{p}^{\mathrm{T}}_{t+j}[\mathtt{none}])\big)\exp(-\bar{\sigma}^{\mathrm{T}}_{t+j}/\rho). \tag{21}$$

This emphasizes teacher-predicted events with lower uncertainty; it does not assume empirically calibrated confidence. The contact loss and the event loss, a Kullback–Leibler (KL) divergence, are

$$\mathscr{L}_{\mathrm{traj}} = \sum_j \omega_j \|\hat{x}^{\mathrm{S}}_{0,t+j} - \mathrm{pack}(\hat{c}^{\mathrm{T}}_{t+j})\|_2^2, \tag{22}$$

$$\mathscr{L}^{\mathrm{S}}_{\mathrm{evt}} = \mathbb{E}[\mathrm{KL}(\hat{p}^{\mathrm{T}} \| \hat{p}^{\mathrm{S}})]. \tag{23}$$

Only trajectory matching uses the event/uncertainty weights. Action distillation compares velocities at the same noisy chunk and noise level:

$$\mathscr{L}_{\mathrm{beh}} = \mathbb{E}\|v^{\mathrm{S}}_{\mathrm{act}}(x_\tau,\tau,o^{\mathrm{S}}) - v^{\mathrm{T}}_{\mathrm{act}}(x_\tau,\tau,o^{\mathrm{T}})\|_2^2. \tag{24}$$

The complete student objective retains reduced-weight ground-truth supervision and the ACC auxiliaries:

$$\begin{aligned}\mathscr{L}^{\mathrm{S}} = {} & \mathscr{L}_{\mathrm{traj}} + \mathscr{L}^{\mathrm{S}}_{\mathrm{evt}} + \mathscr{L}_{\mathrm{beh}} + \mathscr{L}_{\mathrm{ACC}} \\ & + \lambda_{\mathrm{gt}}(\mathscr{L}_{\mathrm{act}} + \mathscr{L}_{\mathrm{vid}} + \mathscr{L}_w) + \lambda'_\sigma \mathscr{L}^{\mathrm{ro}}_{\mathrm{con}}.\end{aligned} \tag{25}$$

The final term (superscript ro: readout) trains the uncertainty readout on detached hidden states. Training mixes demonstrations with recorded teacher and earlier-student rollouts; teacher matching applies even where unsuccessful rollouts have zero ground-truth action weight. We deploy an exponential moving average (EMA) of the student weights, which smooths the parameter trajectory across updates and stabilizes the distilled policy. At inference, the student predicts its own contact package without the teacher or fingertip observations.

Algorithm 1 summarizes training with the teacher and foundation backbone frozen. The deployment intervention in Sec. IV-F changes generated contact, not the training objective.

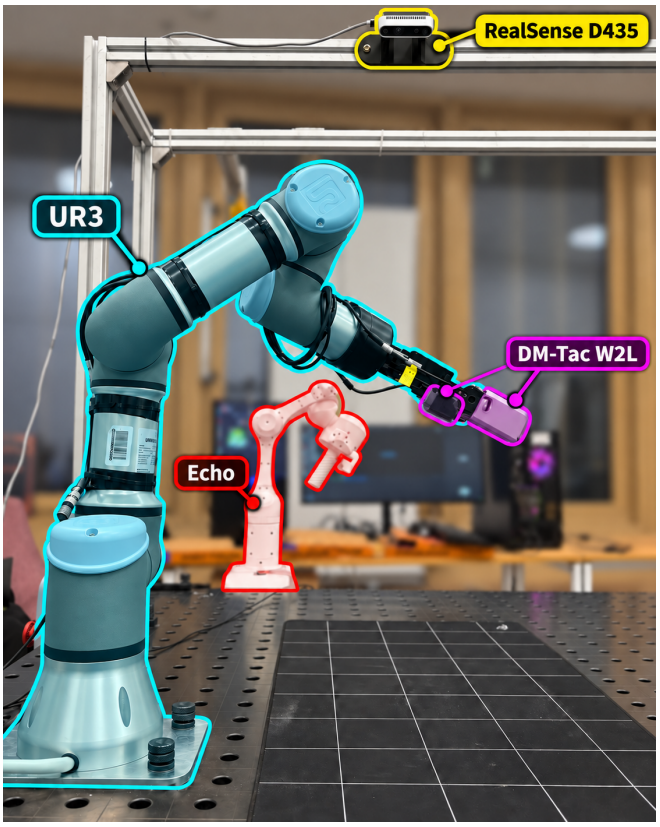

Fig. 4. Our setup consists of a UR3 manipulator with two DM-Tac W2L optical-tactile sensors mounted on the fingers of a Robotiq 2F-85 parallel gripper. We use Realsense D435 to collect scene RGB and Echo to teleoperate the UR3.

IV. Experiments

A. Experimental Setup

We use a UR3 with a Robotiq 2F-85 parallel gripper, two DM-Tac W2L optical fingertips and a static RealSense D435 (Fig. 4). We use Echo teleoperation setup to collect dataset [37]. Each fingertip provides gel images and mechanical fields on a 384-by-288 grid over a 36-by-27 mm surface. Tactile recording is capped at 8 Hz, synchronized robot commands run at 125 Hz and we collect RGB frames from Realsense at 15 Hz.

B. Tasks and Dataset

We evaluate three tasks (Fig. 1):

- **Waffles:** pick the waffles and place them into the box.
- **Carton:** pick the carton and place it into the box.
- **Egg:** pick the egg and place it on the tray.

Each task contributes 250 nominal demonstrations and 35 deliberate-failure demonstrations in which the operator intentionally over- or under-grasps the object, driving the grasp past its feasible force range so that the episode ends in crushing. This gives 285 episodes per task and 855 in total, split into 761 training and 94 validation episodes. Validation episodes never enter training and are drawn only from the nominal set, so no deliberate failure is ever scored. We keep the failure demonstrations because they populate regions of the contact distribution that successful demonstrations never

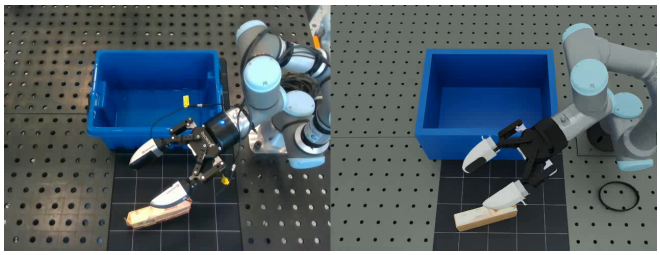

Fig. 5. **(Left)** Real-world setup with UR3, Robotiq gripper and DM-Tac W2L sensors. **(Right)** Simulation setup we created in NVIDIA Isaac Sim.

visit; they are supervised by the contact and video terms with the ground-truth action loss disabled, so they shape what the model predicts about contact without teaching it to reproduce the failing action. Task text accompanies all robot comparisons.

C. Training and Inference

The Cosmos-Predict2.5-2B backbone, VAE and text encoder are frozen. Rank-16 LoRA adapters and new modules give 27.9 M trainable teacher parameters and 26.4 M student parameters. Each training clip contains the current RGB frame and 12 future targets at 4 Hz, spanning 3 s; these are not 13 observed history frames. Temporal compression gives one conditioning video latent and three future latents. Contact targets lie 1, 2 and 3 s after the anchor, while the 16 action targets span 0.1–1.6 s at 10 Hz. At inference only the current image is observed; future video is generated.

Teacher training. We use AdamW [38] in bfloat16 for 20,000 updates on one 24 GB GPU, with effective batch size 8, learning rates of 10^{-4} for LoRA and $3\cdot10^{-4}$ for new modules, 500 warm-up updates, cosine decay and EMA decay 0.999. Contact reconstruction uses a weighting exponent of 0.5. The objective weights, including the ACC gate weight, are

$$(\lambda_a,\lambda_v,\lambda_c)=(1,0.1,1), \tag{26}$$

$$(\lambda_e,\lambda_w,\lambda_g,\lambda_\sigma)=(0.5,0.5,0.2,0.01). \tag{27}$$

Simulation fine-tuning. The deployed teacher receives 2,000 additional updates on 99 scripted simulation episodes and 203 robot rollouts, yielding 2,088 windows. The simulation set includes 98 completed placements and one episode ending with the object still gripped inside the box. We use batch size 8, learning rates reduced fivefold and EMA decay 0.995. Simulation provides contact event labels but no optical tactile measurements; unavailable contact reconstruction, event-band and wrist-wrench losses are masked. Figure 5 shows the corresponding real and simulated scenes.

Student training. The deployed student is initialized from the fine-tuned teacher and distilled for 1,000 updates on one H100, with effective batch size 8. Its dataset contains 860 episodes: the 761 training demonstrations plus 99 earlier rollouts, comprising 66 teacher and 33 student episodes. Eight windows per episode yield 6,880 training windows. Frozen-teacher targets are sampled per batch; failed rollouts receive no ground-truth action weight. Distillation uses

$$(\eta,\rho,\lambda_{\mathrm{gt}},\lambda'_\sigma)=(4,1,0.5,1). \tag{28}$$

TABLE I

DEPLOYED MODEL TRAINING.

Model	Initialized from, then trained
HapticWAM: Teacher	Cosmos-Predict2.5-2B; 20,000 demonstration updates, then 2,000 simulation and rollout updates
HapticWAM: Student	Deployed teacher; 1,000 HID updates on demonstrations and mixed rollouts

Table I summarizes the deployed Teacher-Student pair.

Baselines. We compare against $\pi_{0.5}$ [4] and Diffusion Policy [19], trained on the same 876-episode baseline export, excluding deliberate failures. They receive neither simulation episodes nor robot rollouts. Training uses 20,000 updates at batch size 6 for $\pi_{0.5}$ and 100,000 at batch size 64 for Diffusion Policy (DP). The former updates 430.1 M parameters within a 3.62 B-parameter model; the latter trains all 263.0 M parameters from scratch. DP's deployment history is 0.486 s apart rather than the 0.1 s training interval.

Inference. On an RTX 5090 32 Gb, HapticWAM samples four candidate chunks, screens them using predicted contact/descent and selects by continuity with the previous plan. Median replan times over the 200 trials are 0.445 s for the teacher (compiled kernels), 0.636 s for the student (uncompiled), 0.127 s for $\pi_{0.5}$ and 0.389 s for DP. HapticWAM additionally uses a contact-based closing veto and uncertainty-dependent playback; baseline playback is unscaled.

D. Evaluation Protocol

Each of four models is evaluated on 20 waffle, 20 carton and 10 egg starts, for 200 robot trials. The shared executor retains the same safety guards, but candidate selection, gripper timing and replan limits differ across models. UR3 starting positions are shared across models using seeds 101–120 for waffles/carton and 101–110 for eggs, with model order alternated between launches. Success is an operator-reviewed placement of the object at the target. Independently, every trial receives a sensor-based grasp-quality label from the recorded fingertip force: an over-grasp when the pinch peak exceeds the force reference of Eq. (30), an under-grasp when the fingers contact the object but no held lift follows, and clean otherwise; on eggs, damage is scored by shell cracking. The sensor rule agrees with the operator on placement in 154 of the 160 waffle and carton trials; the operator was not blinded to the running policy. We report the binary placement rate:

$$\text{Success rate}=\frac{100}{N}\sum_{i=1}^{N}\mathbf{1}\{\text{final verdict is placed}\}. \tag{29}$$

Force is reported separately; the placement rate is not redefined by a post-hoc force threshold. Because no independently calibrated damage threshold exists for the waffle and carton, we derive a descriptive force reference from the evaluation trials themselves. Let $\mathscr{P}$ be the set of per-trial pinch peaks, the maximum resultant fingertip force recorded during a trial, over the 73 waffle and carton trials that the sensor rule scores as placements, pooled across all

TABLE II

PICK-AND-PLACE SUCCESS RATE: OPERATOR PLACEMENT VERDICTS.

Model	Waffles ↑	Carton ↑	Egg ↑
Teacher (Ours)	70% (14/20)	60% (12/20)	40% (4/10)
Student (Ours)	**95% (19/20)**	**85% (17/20)**	**50% (5/10)**
$\pi_{0.5}$	20% (4/20)	30% (6/20)	0% (0/10)
Diffusion Policy	10% (2/20)	15% (3/20)	40% (4/10)

four models. With $\mu_{\mathscr{P}}$ and $s_{\mathscr{P}}$ its sample mean and sample standard deviation, the reference is the three-sigma upper envelope of successful-placement forces,

$$F_{\text{ref}} = \mu_{\mathscr{P}} + 3s_{\mathscr{P}} = 24.8\,\text{N}. \tag{30}$$

A pinch peak above F_{ref} is therefore a force atypical of a successful placement of these objects; we call it a reference exceedance rather than damage, since neither object visibly breaks. Eggs are handled differently for two reasons: their damage is directly observable as shell cracking, so no force proxy is needed, and their tolerable force range differs from that of the waffle and carton, so their trials do not enter $\mathscr{P}$ and their over-grasp label is not defined by F_{ref}; egg forces are still reported, but only for description. Reported forces use the resultant-force readout in newtons under the configured sensor scale, not the uncalibrated per-pixel force field.

E. Main Results

Table II compares the deployed teacher, its distilled student, $\pi_{0.5}$ and DP. The student achieves the highest observed placement rate on each task: 95% on waffles, 85% on cartons and 50% on eggs. Paired exact sign tests over shared starts (four-level outcome: no held grasp, held, lifted, placed) put the student above $\pi_{0.5}$ on every task ($p \leq 0.016$) and above DP on waffles and cartons ($p < 0.001$; eggs $p = 0.69$); the student–teacher gap is not separable at this size ($p = 0.062$, 0.125 and 1.0). Across the 50 starts, it places 41 objects, versus 30 for the teacher, 10 for $\pi_{0.5}$ and 9 for DP. The largest student–teacher differences occur on waffles and cartons. Egg handling remains challenging for all methods, and its smaller trial count limits precision: the student's 50% egg success has a descriptive 95% Wilson interval of 23.7–76.3%. Task-wise 95% Wilson intervals are 76.4–99.1% (waffles), 64.0–94.8% (cartons) and 23.7–76.3% (eggs). These observations concern the complete configurations described above; they do not isolate a contact-distillation effect from training-data and execution differences.

F. Contact-Imagination Ablation

Intervention. We evaluate the same deployed student checkpoint without further training, replacing the generated contact frames with a zero contact package throughout sampling. The checkpoint weights and available observation channels are unchanged. The intervention clamps the contact representation, rather than removing a physical sensor or retraining the student. We compare 30 additional robot trials against the intact student's previously recorded outcomes on the same first ten starts per task. Both conditions use the final operator placement verdict.

Results. Clamping contact frames reduces placements on every task (Table IV), from 21 to 4 over the 30 shared starts. The absolute success-rate reduction is 56.7 percentage points, with the largest task-level drop on waffles. Both columns are recorded trials scored by the same operator placement verdict; the intact column is the first ten starts per task of Table II.

Interpretation. The result supports a deployment-time dependence on the contact-generation pathway in the evaluated system, and an offline probe on recorded demonstrations points the same way. On the 94 held-out validation episodes we sample the student as at deployment during the last 1.5 s before the demonstrator first closes the gripper, and measure the *terminal endpoint error*: the distance between the hand position the predicted 1.6 s action chunk would reach and the position the demonstrator actually reached. A policy that does not move scores 28.2 mm. The intact student scores 22.31 mm; with its contact frames clamped to zero it scores 28.31 mm, no better than not moving, whereas zeroing only the ACC input leaves it at 22.35 mm. The harm therefore comes from the content of the generated contact frames rather than from the ACC input, so the comparison is not an isolated test of ACC's attention bias, nor a measure of contact-prediction accuracy. It also does not isolate the mechanism on the robot: clamping changes the jointly generated representation, and contact-dependent selection and closing rules can also mediate the executed outcome. Nor does it establish that contact supervision is better than action-only distillation: both conditions use the same HID-trained weights. The intact trials precede the intervention trials, so matching start seeds does not remove session-order effects.

G. Force Analysis

Table III reports pinch force conditional on sensor-defined placement; these samples can differ from the final operator-reviewed placement counts. The student has lower mean force than the teacher on waffles and eggs, with similar carton forces. Four reference exceedances occurred across the waffle/carton trials: two for the teacher on waffles and two for $\pi_{0.5}$ on cartons. None was recorded for the student. Under-grasps (contact without a held lift): teacher 5/1/2, student 0/2/1, $\pi_{0.5}$ 7/3/4 and DP 10/1/1 on waffles/cartons/eggs; the student alone has neither an over- nor an under-grasp on waffles. The reference is an in-sample waffle/carton band, not calibrated for eggs, where every model exceeds it. Two egg cracks were recorded for DP and none for the other models. These descriptive observations do not establish damage-free control: force summaries exclude failed placements and event detection is limited by sensor coverage.

Force entries show mean and standard deviation where reported; the single-sample and two-sample DP summaries retain only the mean.

TABLE III

PINCH FORCE AT SENSOR-DEFINED PLACEMENTS, MEAN AND STANDARD DEVIATION.

	Waffles		Carton		Egg	
Model	Force (N)	Samples	Force (N)	Samples	Force (N)	Samples
Teacher (Ours)	16.7 ± 4.7	13	11.4 ± 1.4	12	24.9 ± 9.2	5
Student (Ours)	14.3 ± 3.6	19	11.8 ± 2.2	16	18.8 ± 9.1	6
$\pi_{0.5}$	13.4 ± 2.0	5	16.0 ± 4.9	5	—	0
Diffusion Policy	14.1	1	10.1	2	27.1 ± 3.1	4

TABLE IV

CONTACT-IMAGINATION ABLATION: PLACEMENT SUCCESS RATE.

Task	Intact student ↑	Zero contact ↑
Waffles	90% (9/10)	10% (1/10)
Carton	70% (7/10)	30% (3/10)
Egg	50% (5/10)	0% (0/10)
All tasks	70% (21/30)	13.3% (4/30)

H. Student–Teacher Gap

The tactile-free student places more objects than the tactile teacher it was distilled from (41 versus 30 of 50). Three properties of HID make this direction plausible rather than paradoxical.

Corrective on-policy targets. The student's 99 rollout episodes include 33 of its own. On failed rollouts the ground-truth action loss is masked, but Eq. (24) still supplies the frozen teacher's velocity $v_{\text{act}}^{\text{T}}$ at every visited state. This is a DAgger-style expert query [31] with the teacher as expert: the student is corrected on the states it actually reaches at deployment, whereas no comparable expert exists for the teacher, whose rollouts can only be relabeled by outcome. The reversal between offline endpoint error, measured on demonstration states, and closed-loop placement is the expected signature of such a covariate-shift correction.

A tactile-independent contact gate. The student fixes $\alpha_t \equiv 1$ in Eq. (17), so its attention bias depends only on the anticipatory summary,

$$g_t^{\text{S}} = g_t^{\text{ant}} = \text{sigmoid}(W_g \mathbf{u}_t), \tag{31}$$

whereas the teacher's gate also tracks g_t^{react}, a function of tactile-field change sampled at 8 Hz (Sec. IV-A) that includes the uncalibrated per-pixel force field. The student cannot be perturbed by this channel. Consistent with this, on waffles the teacher records five under-grasps and two reference exceedances where the student records none (Sec. IV-G).

Uncertainty-filtered targets. With $(\eta, \rho) = (4, 1)$ from Eq. (28), the per-step weight of Eq. (21) lies in

$$\omega_j \in \left[e^{-\bar{\sigma}_{t+j}^{\text{T}}},\, 5e^{-\bar{\sigma}_{t+j}^{\text{T}}} \right], \tag{32}$$

so a teacher contact prediction at unit uncertainty carries at most 0.37 times the weight of a confident one, and predicted events up to five times the weight of quiescent steps. The student therefore regresses a selectively confident, event-emphasized version of the teacher rather than the teacher itself [39]; with $\lambda_{\text{gt}} = 0.5$ ground-truth anchoring, this can smooth the teacher's errors on precisely the steps where it is unsure.

V. CONCLUSION

HapticWAM combines heterogeneous tactile encoding, structured contact generation and contact-conditioned attention within a world–action model. Haptic-Imagination Distillation transfers contact and action predictions to a student without fingertip tactile inputs. Across the evaluated tasks, the student achieves a 77% per-task mean pick-and-place success rate (82% pooled) versus 57% (60% pooled) for the teacher, and the highest observed placement rate among the evaluated configurations. The common execution controller retains tactile safeguards. Clamping its generated contact frames to zero reduces placement success on shared starts, supporting deployment-time dependence on this pathway. The incremental training benefit of contact supervision over action-only distillation remains to be isolated.

REFERENCES

[1] A. Brohan, N. Brown, J. Carbajal, *et al.*, "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," *arXiv preprint arXiv:2307.15818*, 2023.

[2] M. J. Kim, K. Pertsch, S. Karamcheti, *et al.*, "OpenVLA: An Open-Source Vision-Language-Action Model," in *Conference on Robot Learning (CoRL)*, 2024.

[3] K. Black, N. Brown, D. Driess, *et al.*, "π_0: A Vision-Language-Action Flow Model for General Robot Control," *arXiv preprint arXiv:2410.24164*, 2024.

[4] P. Intelligence, K. Black, N. Brown, *et al.*, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," 2025. [Online]. Available: https://arxiv.org/abs/2504.16054

[5] J. Huang, S. Wang, F. Lin, *et al.*, "Tactile-vla: unlocking vision-language-action model's physical knowledge for tactile generalization," *arXiv preprint arXiv:2507.09160*, 2025.

[6] C. Zhang, P. Hao, X. Cao, *et al.*, "Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation," *Biomimetic Intelligence and Robotics*, p. 100333, 2026.

[7] J. Bi, K. Y. Ma, C. Hao, M. S. Zheng, and H. Soh, "Vla-touch: Enhancing vision-language-action model with dual-level tactile feedback," *IEEE Robotics and Automation Letters*, 2026.

[8] N. Team, F. T. Team, *et al.*, "*n*_0-vtla: Scaling vision-tactile-language-action model with latent tactile tokens," *arXiv preprint arXiv:2607.23782*, 2026.

[9] K. Gubernatorov, M. Sannikov, I. Mikhalchuk, *et al.*, "HapticVLA: Contact-Rich Manipulation via Vision-Language-Action Model without Inference-Time Tactile Sensing," *arXiv preprint arXiv:2603.15257*, 2026.

[10] N. Agarwal, A. Ali, M. Bala, *et al.*, "Cosmos World Foundation Model Platform for Physical AI," *arXiv preprint arXiv:2501.03575*, 2025.

[11] S. Ye, Y. Ge, K. Zheng, *et al.*, "World Action Models are Zero-shot Policies," *arXiv preprint arXiv:2602.15922*, 2026.

[12] R. Li, H. Zhang, J. Jin, *et al.*, "World-Value-Action Model: Implicit Planning for Vision-Language-Action Systems," *arXiv preprint arXiv:2604.14732*, 2026.

[13] P. Tang, S. Xie, B. Huang, *et al.*, "Towards Predictive, Aligned, and Scalable Robot Learning," *arXiv preprint arXiv:2607.11270*, 2026.

[14] T. Yuan, Z. Dong, Y. Liu, and H. Zhao, "Fast-WAM: Do World Action Models Need Test-time Future Imagination?" *arXiv preprint arXiv:2603.16666*, 2026.

[15] Y. Lou, Y. Ye, Y. Fu, *et al.*, "Dream-Tac: A Unified Tactile World Action Model for Contact-Rich Robot Manipulation," *arXiv preprint arXiv:2606.08737*, 2026.

[16] S. Wu, L. You, J. Zhu, *et al.*, "Tactile-WAM: Touch-Aware World Action Model with Tactile Asymmetric Attention," *arXiv preprint arXiv:2606.26663*, 2026.

[17] N. Team, F. T. Team, *et al.*, "*n*_0-twam: Scaling tactile-native world-action model for contact-rich manipulation," *arXiv preprint arXiv:2607.23783*, 2026.

[18] C. Xue, C. Zhang, W. Ma, *et al.*, "Hitac-wam: A hierarchical tactile world action model for contact-rich robot manipulation," *arXiv preprint arXiv:2608.19574*, 2026.

[19] C. Chi, Z. Xu, S. Feng, *et al.*, "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion," in *Robotics: Science and Systems (RSS)*, 2023.

[20] H. Luo, W. Zhang, Y. Feng, *et al.*, "Being-H0.7: A Latent World-Action Model from Egocentric Videos," *arXiv preprint arXiv:2605.00078*, 2026.

[21] S. Tian, Y. Zheng, Y. Zheng, *et al.*, "VT-WAM: Visual-Tactile World Action Model for Contact-Rich Manipulation," *arXiv preprint arXiv:2607.02503*, 2026.

[22] Y. Zang, Y. Zheng, X. Nie, *et al.*, "TacForeSight: Force-Guided Tactile World Model for Contact-Rich Manipulation," *arXiv preprint arXiv:2606.11184*, 2026.

[23] Z. Ma, X. Wei, J. Jiang, S. Lu, and L. Zhang, "Tacpac: Tactile prediction and real-time action correction in world-action models for contact-rich manipulation," *arXiv preprint arXiv:2609.05266*, 2026.

[24] S. Bien, C. Kneissl, T. Jülg, *et al.*, "FlowTouch: View-Invariant Visuo-Tactile Prediction," *arXiv preprint arXiv:2603.08255*, 2026.

[25] C. Higuera, S. Arnaud, B. Boots, *et al.*, "Visuo-Tactile World Models," *arXiv preprint arXiv:2602.06001*, 2026.

[26] G. Ye, Z. Zhang, X. Zhao, *et al.*, "Learning to Feel the Future: DreamTacVLA for Contact-Rich Manipulation," *arXiv preprint arXiv:2512.23864*, 2025.

[27] H. He, Z. Yan, Q. Liu, N. Guo, and W. Lian, "Fawam: Force-aware world action models for closed-loop contact-rich manipulation," *arXiv preprint arXiv:2606.08555*, 2026.

[28] R. Chen, M. Mukadam, M. Kaess, *et al.*, "Ptld: Sim-to-real privileged tactile latent distillation for dexterous manipulation," *arXiv preprint arXiv:2603.04531*, 2026.

[29] R. Zhao, W. Wang, Y. Ma, *et al.*, "FD-VLA: Force-Distilled Vision-Language-Action Model for Contact-Rich Manipulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

[30] Z. Zhang, A. Desai, J.-C. Hu, *et al.*, "Imagining the sense of touch: Touch-informed manipulation via imagined tactile representations," *arXiv preprint arXiv:2607.01684*, 2026.

[31] S. Ross, G. Gordon, and D. Bagnell, "A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[32] W. Yang, W. Liu, R. Xie, *et al.*, "Learning beyond Teacher: Generalized On-Policy Distillation with Reward Extrapolation," *arXiv preprint arXiv:2602.12125*, 2026.

[33] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 4172–4182.

[34] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[35] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[36] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *European conference on computer vision*. Springer, 2016, pp. 47–54.

[37] A. Bazhenov, S. Satsevich, S. Egorov, F. Khabibullin, and D. Tsetserukou, "Echo: An open-source, low-cost teleoperation system with force feedback for dataset collection in robot learning," *arXiv preprint arXiv:2504.07939*, 2025.

[38] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[39] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," in *International conference on machine learning*. PMLR, 2018, pp. 1607–1616.