ζ_0: A Framework for World Action Models with Joint Scene and Body Foresight in Humanoid Control

Yizhi Chen[1,2,*], Zhanxiang Cao[3,2,*], Yitong Wang[3], Junyu Liu[4], Hongchu Zeng[3], Xinyi Peng[5,2], Chuxuan Huang[6,2], Yicheng Su[6], Zihao Zhou[3,2], Shengcheng Fu[1,2], Haoyang Li[3,2], Liyun Yan[3,2], Tianyue Zhan[3,2], Jianming Ma[3,2], Yang Zhang[3], Keqi Zhu[5,2], Xiaxi Si[3], Jinming Yao[7], Qiao Sun[8], Shizeng Zhang[8], Jiangwei Zhong[8], Yue Gao[3,2,†]

[1]Tongji University [2]Shanghai Innovation Institute [3]Shanghai Jiao Tong University [4]Shanghai University
[5]Zhejiang University [6]Fudan University [7]University of Science and Technology of China [8]Lenovo Group Limited

[*]Equal contribution. [†]Corresponding author.

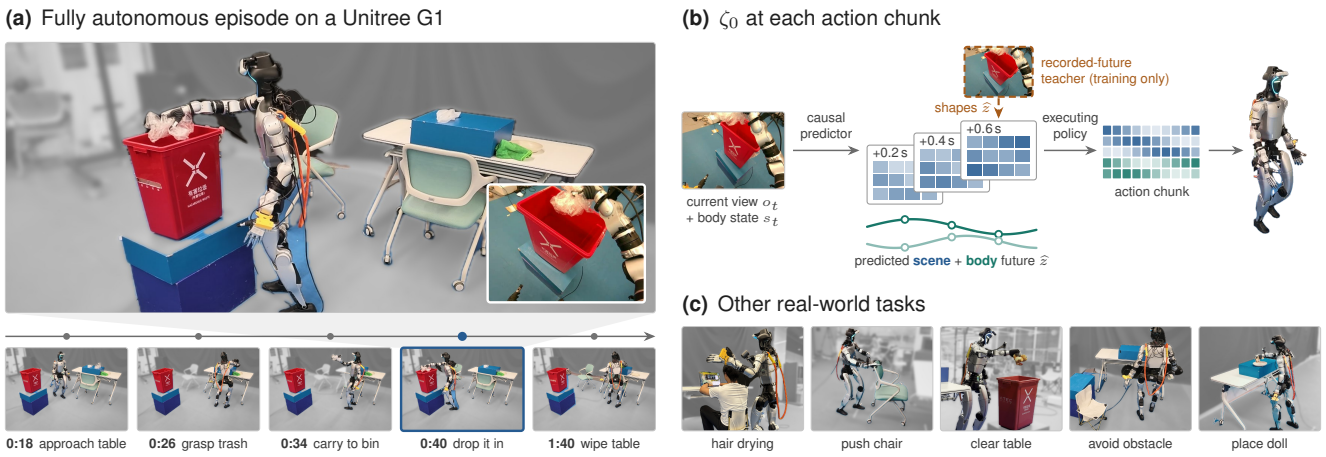


Figure 1. **ζ_0, a framework for World Action Models with joint scene and body foresight.** (a) An autonomous ζ_0 (GR00T) episode without human intervention: walking to the table, carrying trash to the bin, and wiping the table. Five successive moments are shown; the enlarged 0:40 frame includes the synchronized head-camera view. Third-person backgrounds are desaturated and blurred. (b) Causally predicted scene-and-body futures condition action generation; a teacher conditioned on recorded futures supervises prediction through action responses during training only. (c) Additional real-world tasks, including interaction with a person.

Abstract

World Action Models anticipate visual outcomes to guide robot actions, yet humanoid interaction also depends on how the robot's body responds. Visual foresight alone can miss body motion, while latent-distance losses do not directly supervise action responses. We present ζ_0, a framework that turns vision-language-action policies into World Action Models by foreseeing future visual features and measured joint trajectories in a shared latent representation. During training, a teacher conditioned on recorded futures provides a reference action response, and dual-response transfer trains the causal predictor to match it through the teacher and the executing policy. All components are trained jointly, and deployment uses only causally predicted futures. On seven HumanoidArena tasks, ζ_0 built on Ψ_0 and GR00T N1.7 reaches 63.6% and 79.5% mean success, 18.8 and 9.0 points above these bases. On six real-world tasks, it reaches 78.3% and 85.0%, gains of 16.7 and 10.0 points. On GR00T, its 79.5% tops all compared seven-task joint policies. Project page: https://chenyizhi123.github.io/zeta0-project/.

1. Introduction

Humanoid interaction couples whole-body motion with changes in the surrounding scene. Pushing a box, for example, moves the object while the arms keep contact and the legs and torso support the motion. World Action Models couple action generation with anticipated future observations [28, 33, 46, 50, 51, 56]. For a humanoid, the consequences of an action also include its own body motion, which egocentric images may capture only partially and measured joint trajectories record directly. We therefore ask a policy to foresee scene and body changes jointly.

Demonstrations provide this foresight as supervision: each action is followed by recorded images and measured body states. At execution time, those futures have not yet occurred and must be predicted from causal inputs. A common approach fits future predictions to recorded observations or their latent representations [21, 23, 56]. Latent-distance supervision, however, does not directly reflect how a policy uses the prediction: nearby futures need not induce the same action, and equally sized errors can have different effects on behavior. We therefore also supervise prediction in action

space, asking whether a predicted future induces the action that the recorded future would.

We introduce ζ_0, a framework that turns a vision-language-action policy into a World Action Model with joint scene-and-body foresight for humanoid control, illustrated in Fig. 1. During training, a future encoder summarizes subsequent visual features and measured body trajectories into a shared temporal latent representation, and a causal predictor learns to estimate it from inputs available at execution time. The representation, predictor, and action models are trained jointly; at deployment, the policy acts on the predicted future and current observations.

To supervise prediction in action space, we introduce *dual-response transfer*. A teacher learns to generate actions conditioned on recorded futures, providing a reference action response, and we train the predictor so that predicted futures induce this response through two readers. The teacher term asks the prediction to carry the action-relevant information the teacher extracts from recorded futures; the executing-policy term asks for the same response from the reader used at deployment.

We make three contributions:

- **Joint scene-and-body foresight framework**: ζ_0 turns vision-language-action policies into World Action Models.
- **Dual-response transfer**: a training objective that supervises causal prediction through the responses of a recorded-future teacher and the executing policy.
- **Consistent gains in simulation and the real world**: ζ_0 built on Ψ_0 and GR00T N1.7 raises mean success over each base by 18.8 and 9.0 points on seven HumanoidArena tasks and by 16.7 and 10.0 points on six real-world tasks.

2. Related Work

Whole-body humanoid learning. Whole-body humanoid learning studies coordinated locomotion, posture, and manipulation, with hierarchical control separating high-level decisions from low-level execution [9, 13]. Tracking controllers such as ExBody, ExBody2, HOVER, and SONIC learn to execute reference motions or end-effector targets across the body [14, 26, 30, 39], and ASAP aligns simulated and real-world physics to learn agile whole-body skills [25]. Teleoperation systems collect coordinated demonstrations [24, 53], while HumanPlus and 3D diffusion policies connect visual observations to autonomous humanoid skills [17, 54]. HumanoidBench evaluates locomotion and manipulation across simulated tasks [45], and HumanoidArena emphasizes egocentric hierarchical whole-body learning [47]. These directions provide motion-control interfaces and demonstration-based learning settings for humanoid interaction; ζ_0 builds on such an interface and asks how a high-level policy can anticipate the consequences of its own commands.

Vision-language-action policies. Recent VLA work adapts pretrained vision-language representations to robot control using demonstration data [6, 11, 31]. Building on large-scale robot transformers [7], generalist policies such as Octo and the π series learn across tasks and embodiments [4, 42, 44]. Action generation has developed alongside representation learning, with action chunking [55] and diffusion- or flow-based generators [4, 15] providing interfaces for continuous robot commands. Receding-horizon execution refreshes the observation context between action chunks, combining multi-step generation with feedback during control [15]. For humanoids, GR00T N1, Ψ_0, and HuGe combine pretrained vision-language backbones with action generators [38, 41, 49], and Kimodo-Policy learns kinematic action models for loco-manipulation [48]. These policies provide foundations for studying how additional predictive information can condition action generation.

World Action Models. World models learn how an environment responds to actions, either to learn behaviors by imagining latent trajectories [21–23] or as generative and predictive models trained on large video corpora [2, 8, 40]. Text-guided video generation can serve directly as a policy [16], and frozen self-supervised visual features can serve as world-model states for planning [57]. In feature-space planning, candidate action sequences are optimized so that predicted future features approach those of a goal observation [57]. Predictive policies couple action learning with future visual prediction [28, 46, 50] or jointly model future observations and actions [34, 58]. FLARE aligns policy features with future visual latents [56], DIAL separates latent intent prediction from execution [12], and ARPS learns predictive states from future features and their changes [20]. Being-M0.7 jointly models visual and motion futures, then trains an action expert on a frozen, robot-adapted prior [52]. Most of these methods represent the future by images or visual features and supervise prediction by reconstruction or latent distance. ζ_0 foresees the scene and measured body trajectories in one shared latent representation and supervises its causal prediction in action space, keeping the predictor coupled to action learning throughout training.

Privileged information and action supervision. Learning with privileged information transfers guidance from training-only observations to deployable policies [10, 32]. Knowledge distillation matches a student's outputs to a teacher's and thereby updates the student's parameters [27]. Action-space constraints provide another form of supervision: cross-view action consistency matches flow velocities across action-equivalent observations [29]. Dual-response transfer in Sec. 3.5 also compares action responses, but it holds the readers' parameters fixed and updates the predictor only through the readers' input, the predicted future.

3. Method

ζ_0 is a framework that turns a base vision-language-action policy into a World Action Model by adding joint scene-and-body foresight, a causal prediction of how the scene and the robot's body will change over a short future horizon. The base supplies the vision-language backbone, the state and action interfaces, and, where available, pretrained weights. It adds a shared latent representation of future scene and body changes, predicts it from causal inputs, and conditions action generation on the prediction. We instantiate ζ_0 on Ψ_0 and GR00T N1.7, denoting the resulting models by ζ_0 (Ψ_0) and ζ_0 (GR00T), as set up in Sec. 4.1. Sections 3.2–3.6 describe the representation, its causal and world prediction, action learning, dual-response transfer, and joint optimization and deployment; Fig. 2 separates training from deployment.

3.1. Problem formulation

At time t, let $x_t = (o_t, s_t, \ell_t)$ contain the current image observation, measured state, and instruction. Let p_t be the prefix of an action chunk already committed by the controller [5]. A training demonstration additionally provides an action chunk a_t and a continuation y_t of future visual features and body measurements, with timestamps and validity masks. We divide this continuation into K temporal groups ending at increasing horizons. Each group contains a visual target and body measurements at timestamped query points; the body target is a trajectory rather than a single joint vector at the group endpoint.

The base policy's vision-language backbone extracts current visual-language context and visual features for future supervision. Although both may use joint coordinates, the measured body target differs from the commanded action: the action is a target for the low-level controller, whereas the measured trajectory records how the body actually moves under contact, balance, and tracking dynamics, which is the consequence the policy must anticipate.

3.2. Learning scene-and-body futures

An online encoder E_ϕ maps the current visual and body context x_t^{vb} and the recorded continuation to $z_t \in \mathbb{R}^{K\times N\times d}$, with N tokens of width d per group:

$$z_t = E_\phi(x_t^{\mathrm{vb}}, y_t), \qquad z_t^\star = \mathrm{sg}\left(E_{\bar\phi}(x_t^{\mathrm{vb}}, y_t)\right). \tag{1}$$

The target parameters $\bar\phi$ track ϕ by exponential moving average (EMA), as in self-supervised predictive learning [1, 19]. Neither encoder reads the instruction or demonstrated actions; both summarize only the recorded consequences. The teacher defined below therefore receives privileged information that is available only in training [10, 32]; its action loss updates the online encoder to make these consequences useful for action generation.

A readout D takes the latent and current visual features and reconstructs future scene features and measured body trajectories. Write $\ell_s(z; D)$ and $\ell_b(z; D)$ for their masked mean squared errors. The temporal term $\ell_\Delta(z; D)$ sums scene and body errors on differences between adjacent groups, using only valid corresponding endpoints. We define

$$\mathcal{L}_{\mathrm{rec}}(z; D) = \ell_s(z; D) + \ell_b(z; D) + \lambda_\Delta \ell_\Delta(z; D), \tag{2}$$

where λ_Δ weights the temporal term. This supervision preserves both the state of the interaction and its temporal change.

3.3. Shared causal and world prediction

A conditioner H encodes the causal context as $h_t = H(x_t)$. A shared deterministic predictor F_θ predicts both modes:

$$\begin{aligned} \hat z_t^{\mathrm{prior}} &= F_\theta(h_t, p_t; \mathrm{prior}), \\ \hat z_t^{\mathrm{world}} &= F_\theta(h_t, a_t; \mathrm{world}). \end{aligned} \tag{3}$$

The prior mode uses the committed prefix; the world mode additionally uses demonstrated future actions. Both modes share parameters and predict all temporal groups in parallel. The prior mode predicts what will happen from information available at decision time and is the mode used at deployment; the world mode predicts what happens under given future actions, an action-conditioned dynamics problem with less ambiguity. Sharing parameters lets the prior predictor reuse the dynamics learned in world mode, and Sec. 4.3 shows that removing the world latent MSE lowers success on both bases.

For mode m, we minimize

$$\mathcal{L}_{\mathrm{pred}}^m = \lambda_m \|\hat z_t^m - z_t^\star\|_{M_m}^2 + \lambda_D \mathcal{L}_{\mathrm{rec}}(\hat z_t^m; D_{\mathrm{fix}}), \tag{4}$$

where λ_m and λ_D are loss weights and $\|\cdot\|_{M_m}^2$ is a masked mean squared error. The readout parameters are detached for this prediction loss, while derivatives with respect to its latent input remain active. Thus D provides a common measurement of recorded and predicted futures. It is trained separately by Eq. (2) on online-encoder latents.

3.4. Action learning from recorded and predicted futures

We use distinct flow-matching [36, 37] action models: teacher T_ψ and executing policy G_ω. For a normalized demonstration $\tilde a$, Gaussian noise ϵ, and a sampled flow time $\tau \in [0, 1]$, the uncommitted action suffix is

$$a_\tau = (1-\tau)\tilde a + \tau\epsilon, \qquad v^\star = \epsilon - \tilde a. \tag{5}$$

Committed prefix entries remain clean, have flow time zero, and are excluded from the loss. Let $X = (a_\tau, \tau, x_t, p_t)$, and let M_a select valid, uncommitted action entries. The flow-matching action objectives are

$$\begin{aligned} \mathcal{L}_T &= \mathbb{E}\|T_\psi(X, z_t) - v^\star\|_{M_a}^2, \\ \mathcal{L}_G &= \mathbb{E}\|G_\omega(X, \hat z_t^{\mathrm{prior}}) - v^\star\|_{M_a}^2. \end{aligned} \tag{6}$$

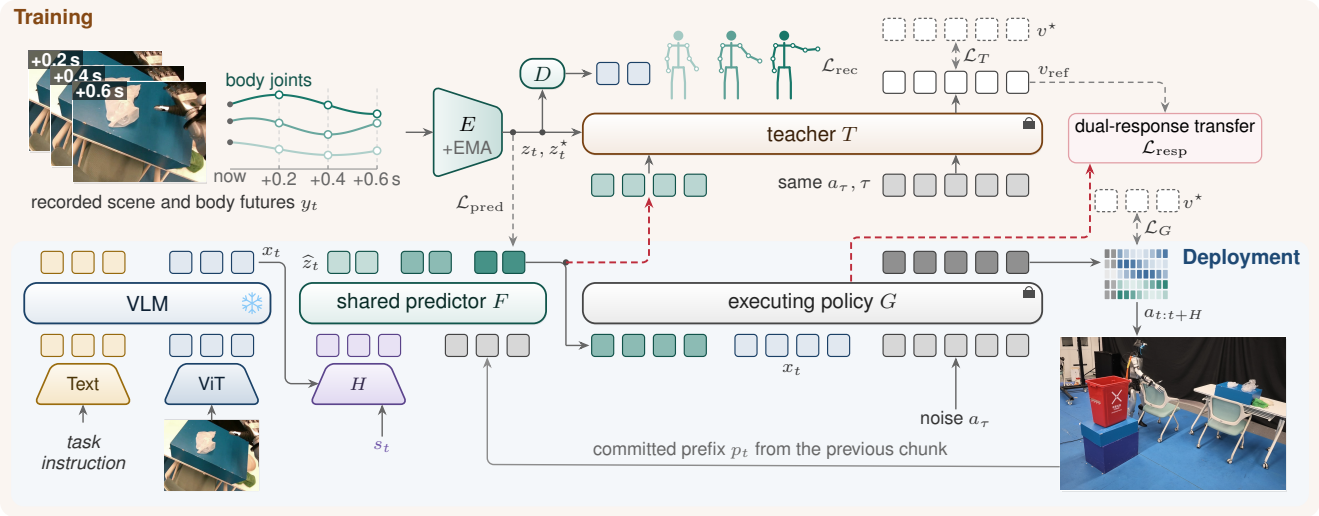


Figure 2. **Overview of ζ_0.** *Bottom, deployed:* the base VLM encodes the instruction and current view; with the measured state, the conditioner H produces h_t, from which the shared predictor F and the committed prefix p_t predict a latent future $\widehat{z}_t$ over the temporal groups. The flow-matching policy G generates the action chunk from $\widehat{z}_t$, the current context x_t, and noisy actions. *Top, training only:* the encoder E and its EMA target summarize the recorded future into z_t and $z_t^\star$; the readout D reconstructs scene features and body trajectories, the teacher T learns actions from z_t, and $z_t^\star$ supervises both the prior and, in dashed gray, the action-conditioned world mode of F. In dual-response transfer, drawn in red, T and G read $\widehat{z}_t$ on the same flow sample with their parameters locked and match the recorded-future reference v_{ref}; this gradient reaches only $\widehat{z}_t$, F, and H. Gray dashed $\mathcal{L}_T$ and $\mathcal{L}_G$ are the flow-matching action losses that train T and G against the demonstration flow target $v^\star$, drawn as dashed tokens, in Eq. (6).

The teacher loss trains T and the online future encoder. The executing-policy loss trains G and backpropagates through the causal prediction path. Both action models, which we call readers of the future latent, have direct access to current visual-language context and measured state.

3.5. Dual-response transfer

The teacher learns to generate actions from recorded futures. Dual-response transfer uses this learned action mapping to supervise causal prediction, asking predicted futures to induce the teacher's recorded-future response. For one shared flow sample X, the teacher conditioned on the EMA future defines

$$v_{\text{ref}} = \text{sg}\left(T_\psi(X, z_t^\star)\right). \tag{7}$$

Writing $\widehat{z} = \widehat{z}_t^{\text{prior}}$, we apply

$$\begin{aligned}\mathcal{L}_{\text{resp}} = \tfrac{1}{2}\|T_{\text{fix}}(X, \widehat{z}) - v_{\text{ref}}\|^2_{M_r} \\ + \tfrac{1}{2}\|G_{\text{fix}}(X, \widehat{z}) - v_{\text{ref}}\|^2_{M_r},\end{aligned} \tag{8}$$

where M_r combines valid action-suffix entries with availability of future supervision. The subscript "fix" detaches reader parameters and buffers within this objective; it does not detach $\widehat{z}$. All other response inputs and the reference are detached, so this loss updates the prediction path through $\widehat{z}$.

The teacher term asks the same reader to interpret the prediction as it interprets the recorded future. The executing-policy term asks for that response in the reader used at deployment. These are different constraints: even if $\widehat{z} = z_t^\star$, the second residual need not vanish because G and T are different models. The objective thus supervises prediction in action space, complementing the latent MSE of Eq. (4) rather than measuring latent prediction error alone. The flow-matching action objectives in Eq. (6) continue to update both readers; "fix" applies only within response transfer.

3.6. Joint optimization and deployment

During joint learning, forecasting alternates between prior and world updates. The executing policy always trains on a causal prior prediction, including on world-forecast updates. Each update uses

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{rec}}(z_t; D) + \mathcal{L}_T + \lambda_{\text{reg}}\mathcal{R}(z_t) + \mathcal{L}_G \\ + \mathcal{L}^m_{\text{pred}} + \mathbf{1}[m = \text{prior}]\lambda_r\mathcal{L}_{\text{resp}},\end{aligned} \tag{9}$$

where λ_r and λ_{reg} are loss weights. The regularizer bounds latent scale and penalizes collapse of tokens within a group. The vision-language backbone is frozen. Appendix A specifies initialization, masking rules, and parameter updates.

At deployment, we compute $F(H(x_t), p_t; \text{prior})$ once per action chunk, then integrate G from noise at $\tau = 1$ to actions at $\tau = 0$, keeping the committed prefix fixed. The predictor is deterministic; stochasticity enters action sampling. The future encoder, EMA encoder, readout, teacher, and world-mode action labels are absent from deployment.

Table 1. **Seven-task joint-policy success rates (%) on HumanoidArena.** All policies are trained jointly on the seven tasks. Published rows retain source-specific settings; ‡ marks averages over base and three OOD modes from [35], and $*$ denotes additional policy pretraining. Our runs (final six rows) use 700 demonstrations, 40,000 updates, and 60 evaluation episodes per task (Sec. 4.1). Bold marks ζ_0 models.

Policy	Double desk	Football	Box	Boxing	Open door	Sit sofa	Navigation	Mean
HuGe-OFT [38]	0.00	48.33	53.33	73.33	75.00	55.00	73.33	54.05
HuGe-GR00T [38]	15.00	71.67	81.67	68.33	95.00	78.33	56.67	66.67
HuGe-Pi [38]	11.67	63.33	76.67	68.33	88.33	83.33	63.33	65.00
Kimodo-Policy [48]	40.0	35.0	80.0	66.7	93.3	96.7	61.7	67.63
Kimodo-Policy* [48]	38.3	40.0	90.0	68.3	95.0	96.7	76.7	72.14
Fast-WAM‡ [51]	45.00	46.25	73.75	67.92	85.42	74.58	35.00	61.13
LingBot-VA‡ [33]	41.25	51.67	77.92	72.92	84.58	77.08	52.08	65.36
GigaWorld-Policy 0.5‡ [18]	52.50	62.08	65.83	73.33	90.83	74.17	60.00	68.39
ACT [55]	11.67	8.33	76.67	73.33	73.33	73.33	48.33	52.14
$\pi_{0.5}$ [44]	11.67	5.00	80.00	68.33	75.00	66.67	50.00	50.95
Ψ_0 [49]	6.67	26.67	61.67	50.00	65.00	63.33	40.00	44.76
ζ_0 (Ψ_0)	21.67	23.33	85.00	61.67	95.00	96.67	61.67	**63.57**
GR00T N1.7 [41]	43.33	38.33	86.67	75.00	98.33	93.33	58.33	70.48
ζ_0 (GR00T)	45.00	56.67	96.67	88.33	96.67	95.00	78.33	**79.52**

4. Experiments

We evaluate ζ_0 built on Ψ_0 [49] and on GR00T N1.7, a later release of GR00T N1 [41]. On seven HumanoidArena [47] tasks, we compare ζ_0 on each base with that base policy and include ACT [55] and $\pi_{0.5}$ [44] as additional baselines, supplemented by published seven-task joint-policy results. Component ablations on both bases examine the shared future representation and its coupling to action generation. We then measure body forecasts with ζ_0 (Ψ_0) and evaluate both ζ_0 models on six real-world tasks.

4.1. Experimental setup

Benchmark and data. We use the Unitree G1 setup in HumanoidArena and seven tasks: double desk, football, pick and place box, boxing, open door, sit sofa, and visual navigation. Each task contributes 100 demonstrations, giving 700 demonstrations in total. Each policy trained in our experiments is one shared model trained on all seven tasks with uniform task sampling and global batch size 512. In these runs, policies consume current egocentric visual context, a 64-dimensional state, an instruction, and the committed action prefix, and output 40-dimensional semantic actions. Future-body supervision in our models comprises 29 measured joints: 14 arm, 12 leg, and three waist coordinates.

Instantiations. ζ_0 (Ψ_0) uses the frozen Qwen3-VL-2B vision-language backbone [3] released with Ψ_0 and randomly initializes the ζ_0 task modules, whereas the Ψ_0 baseline starts from its full pretrained checkpoint. ζ_0 (GR00T) and the GR00T N1.7 baseline both start from the pretrained GR00T N1.7 checkpoint; in ζ_0 (GR00T), the teacher and executing policy are copies of its native diffusion-transformer [43] action head, and the predicted future enters its cross-attention as additional context tokens. All seven-task policies we train, including ablations and baselines, use 40,000 updates and are evaluated at the final checkpoint; $\pi_{0.5}$ and Ψ_0 start from their pretrained checkpoints. In all experiments, execution uses a 30-step planning window at 50 Hz: the controller executes 24 steps and retains the last six as the committed prefix of the next chunk, corresponding to 0.6, 0.48, and 0.12 s. The forecast covers this window in $K = 3$ groups ending at 0.2, 0.4, and 0.6 s, with $z_t \in \mathbb{R}^{3\times8\times256}$. Loss weights are $\lambda_\Delta = \lambda_D = 0.25$, $\lambda_{\text{prior}} = \lambda_{\text{world}} = \lambda_r = 1$, and $\lambda_{\text{reg}} = 10^{-3}$. Optimizer and latent settings are listed in Appendix A.

Closed-loop evaluation. We use scenario seeds 0, 1, and 2 with 20 episodes per seed, giving 60 episodes per task and 420 per setting. All policies in our runs share the benchmark environments, success predicates, episode limits, and SONIC low-level controller [39], and chunked execution with a committed action prefix, detailed in Appendix A.5. We report per-task success and the equal-weight mean. Scenario seeds vary evaluation conditions, and each of our settings uses one training seed.

4.2. Enhancing different base policies

Table 1 shows that ζ_0 improves both bases. ζ_0 (Ψ_0) reaches 63.6% mean success versus 44.8% for Ψ_0, an 18.8-point gain; it improves six of seven tasks, most on sit sofa, open door, and pick and place box, by 33.3, 30.0, and 23.3 points. ζ_0 (GR00T) reaches 79.5% versus 70.5% for GR00T N1.7, a 9.0-point gain; it improves six of seven tasks, most on

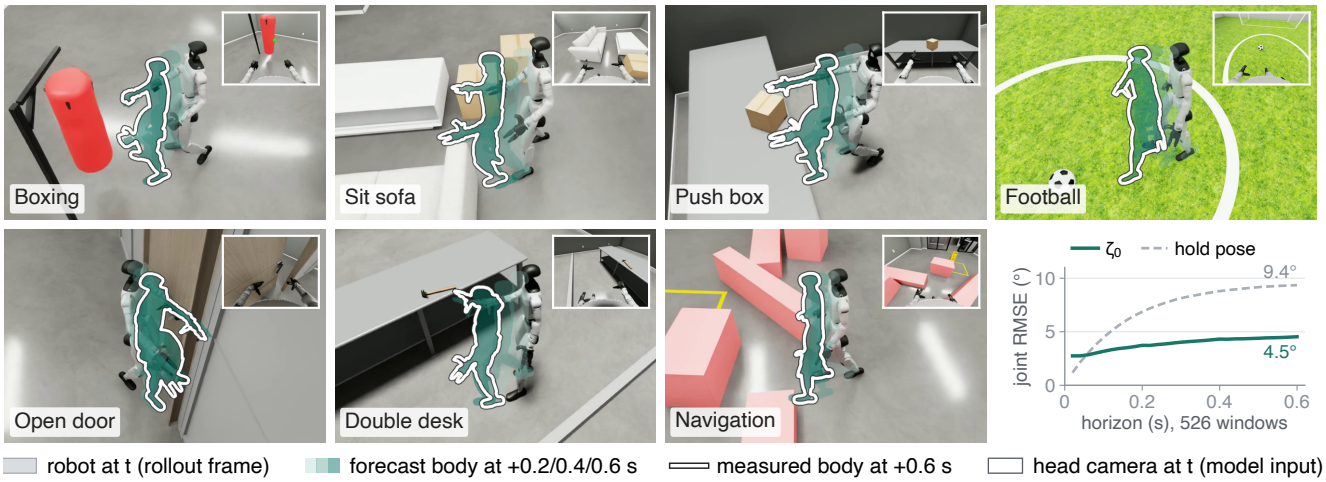


Figure 3. **Forecasting the body in the scene with ζ_0 (Ψ_0).** Each scene panel is a frame at time t from a closed-loop evaluation rollout of the shared seven-task model. From the head-camera view at t, shown as an inset, and the current state, its causal predictor and readout forecast the 29 joint angles over the next 0.6 s; we draw the forecast body at $+0.2$, $+0.4$, and $+0.6$ s in green, darker for later times, through the recorded camera and outline in white the body the robot actually reached at $+0.6$ s. The model does not forecast root motion, so forecast and measured bodies are placed at the measured root pose of their time, and the 14 unpredicted finger joints are held at t. Examples are chosen for clear task context; the last panel reports joint RMSE against horizon over all 526 windows of Sec. 4.4, compared with holding the pose at t.

navigation, football, and boxing, by 20.0, 18.3, and 13.3 points, while open door stays near ceiling at 96.7%. This is the highest reported seven-task mean in Table 1. ζ_0 (GR00T) also has the highest success of any policy in Table 1 on pick and place box, boxing, and navigation, at 96.7, 88.3, and 78.3%. ζ_0 (Ψ_0) exceeds ACT and $\pi_{0.5}$ by 11.4 and 12.6 points and outperforms both on six of seven tasks.

Table 2. **Component ablations of ζ_0 on both bases.** Seven-task mean success (%) over 420 episodes; Full denotes the complete ζ_0 model on each base, and Δ is the change from it in percentage points. Each ablation removes BF, future body measurements as encoder input and supervision, leaving visual futures only; DR, dual-response transfer; ER or TR, the executing-policy or teacher response term; AG, the action-loss gradient from G into H/F; FC, future conditioning of G; or WL, the world latent MSE.

	ζ_0 (Ψ_0)		ζ_0 (GR00T)	
Setting	Mean	Δ	Mean	Δ
Full	**63.57**	—	**79.52**	—
w/o BF	61.43	−2.14	73.10	−6.43
w/o DR	58.10	−5.48	75.00	−4.52
w/o ER	61.67	−1.90	77.14	−2.38
w/o TR	60.95	−2.62	75.24	−4.29
w/o AG	59.52	−4.05	76.90	−2.62
w/o FC	47.14	−16.43	73.33	−6.19
w/o WL	61.90	−1.67	74.76	−4.76
Base policy	44.76	−18.81	70.48	−9.05

4.3. Component analysis

Table 2 removes one component at a time on both bases: future body measurements (BF), dual-response transfer (DR) and its executing-policy and teacher terms (ER, TR), the action gradient into prediction (AG), future conditioning of G (FC), and the world latent MSE (WL). Every ablation lowers the seven-task mean on both bases; per-task results are in Appendix C.

Acting on predicted futures. Zeroing G's future condition in training and deployment lowers the mean by 16.4 points on Ψ_0 and 6.2 on GR00T N1.7, confirming on both bases that the executing policy acts on its foresight. On Ψ_0 it lowers success on six of seven tasks, most on pick and place box, by 23.3 points.

Body futures. Keeping only visual futures, without future body inputs or supervision, lowers the mean by 6.4 points on GR00T N1.7, the largest ablation drop on this base, and by 2.1 on Ψ_0. On GR00T N1.7, the drop concentrates on tasks with object contact and whole-body coordination: football, pick and place box, boxing, and double desk each lose 10.0–13.3 points, whereas open door and sit sofa do not decrease. In these tasks the body's response to contact can be difficult to read from egocentric images.

The effects of body supervision and future conditioning also differ by task. On GR00T N1.7, removing body futures lowers box success by 11.7 points, compared with 1.7 points

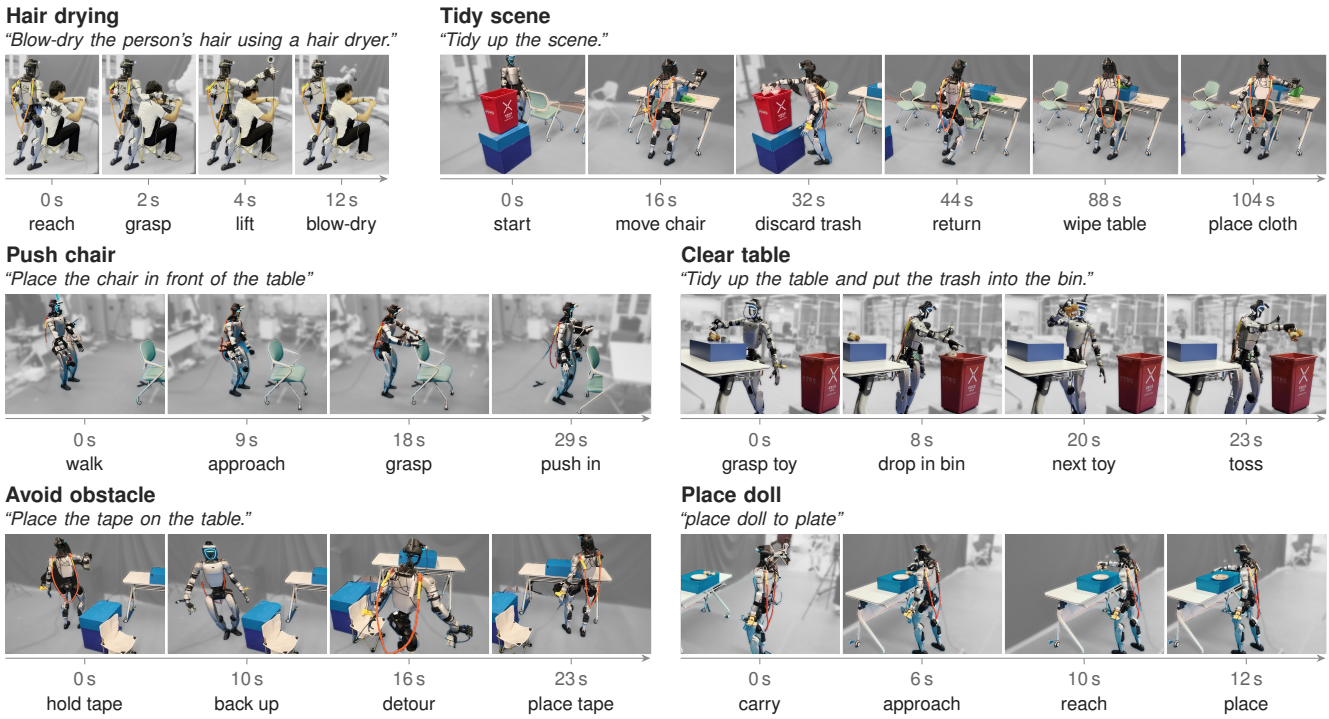


Figure 4. **Real-world tasks on a Unitree G1.** Keyframes from ζ_0 (GR00T) episodes on the six tasks, each titled with its language instruction; times are relative to the first frame shown.

when future conditioning is removed; on navigation, the corresponding drops are 5.0 and 20.0 points. Body supervision is more influential on box, while access to the predicted latent is more influential on navigation.

Dual-response transfer. Removing dual-response transfer lowers the mean by 5.5 points on Ψ_0 and 4.5 on GR00T N1.7; keeping only one reader at the same total coefficient leaves the mean 1.9–4.3 points below the full model. Removing the teacher term lowers the mean by 2.6 and 4.3 points on Ψ_0 and GR00T N1.7, more than the 1.9 and 2.4 points for the executing-policy term, and removing both costs more than removing either, suggesting that the two readers supply partly complementary supervision.

The single-reader controls have different task profiles despite the same total response weight. For ζ_0 (Ψ_0), teacher-only supervision obtains 88.3% on box and 55.0% on boxing, compared with 71.7% and 58.3% for executing-policy-only supervision. On GR00T N1.7, the corresponding boxing rates are 80.0% and 85.0%. Teacher-only supervision has the higher mean on both bases, while executing-policy-only supervision performs better on boxing.

Other components. Blocking the executing-policy action gradient into prediction lowers the mean by 4.0 and 2.6 points on Ψ_0 and GR00T N1.7, and removing the world latent MSE by 1.7 and 4.8 points.

The per-task results in Appendix C show where these losses differ. On ζ_0 (Ψ_0), blocking the action-loss gradient reduces box and boxing success by 21.7 and 16.7 points; removing the world latent MSE lowers sit-sofa success by 16.7 points. On GR00T N1.7, removing the world latent MSE reduces football, box, and navigation by 13.3, 10.0, and 11.7 points, whereas blocking the action-loss gradient mainly affects navigation, by 11.7 points, and leaves boxing, open door, and sit sofa unchanged.

4.4. Forecasting the body

We next measure how well ζ_0 (Ψ_0) anticipates its own body. For each task we record one new closed-loop rollout and place an anchor every 0.1 s over its first 8 s, giving 526 windows; the boxing rollout ends before 8 s. At each anchor, the causal predictor and readout forecast the 29 joint angles over the next 0.6 s from the current observation alone and we compare them with the joint angles the robot reached in that rollout. Over all windows, the joint RMSE is 3.7°, 4.3°, and 4.5° at +0.2, +0.4, and +0.6 s, roughly half the 6.9°, 8.8°, and 9.4° of holding the current pose, as plotted in Fig. 3. From +0.2 to +0.6 s the forecast error grows by only 0.8°, whereas holding the pose grows by 2.5°, so the advantage widens with the horizon. At +0.6 s the forecast is closer than this baseline in 83% of windows, and its error averaged over the 0.6 s horizon is lower on every task. Holding the current

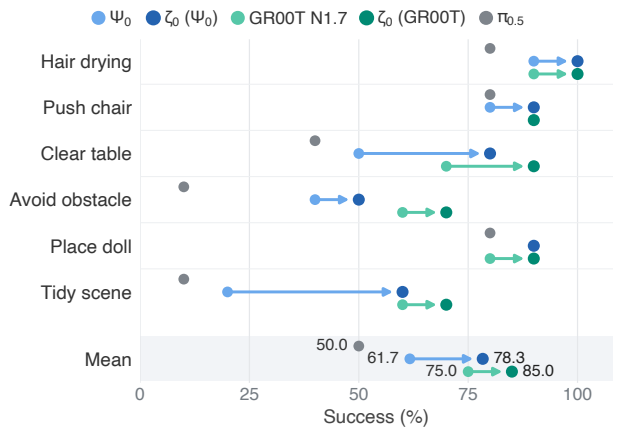


Figure 5. **Real-world success rates on a Unitree G1.** Arrows point from each light base policy to its dark ζ_0 counterpart, in blue for Ψ_0 and teal for GR00T; a single dot marks equal success. Each policy is evaluated for 10 trials per task, 300 in total; Mean weights the six tasks equally. Task criteria are in Appendix B.

pose is hard to beat only within the first 0.05 s; beyond that the forecast is more accurate at every horizon. Figure 3 renders forecasts in their scenes, placing each forecast body at the measured root pose.

Pooling all forecast times within each task, navigation has a trajectory RMSE of 5.2° for prediction versus 11.2° for holding the current pose. The corresponding errors are 3.3° versus 8.0° on sit sofa and 4.1° versus 5.4° on box. Navigation has both the largest forecast error and the largest absolute reduction over pose persistence among the seven rollouts.

4.5. Real-world experiments

Setup. We compare ζ_0 (Ψ_0) and ζ_0 (GR00T) with their respective bases and $\pi_{0.5}$ on six Unitree G1 tasks: hair drying, pushing a chair, clearing a table, carrying tape around an obstacle, placing a doll, and tidying a scene, shown in Fig. 4. The tasks place different demands on the policy: hair drying requires sustained arm motion near a person; pushing a chair and placing a doll combine walking with grasping; carrying tape around an obstacle couples carrying with obstacle-aware locomotion; and table clearing and scene tidying chain several object interactions in one episode. All five policies share three RGB views, measured body-and-hand state, the action interface, the SONIC controller, and the execution protocol, and each is trained separately on each task with the same demonstrations and 20,000 updates. Each policy is evaluated for 10 trials per task with randomized robot and object positions; failures and timeouts count as unsuccessful, and Appendix B lists demonstrations, success criteria, and per-task counts.

Results. ζ_0 (Ψ_0) reaches 78.3% mean success versus 61.7% for Ψ_0 and 50.0% for $\pi_{0.5}$, and ζ_0 (GR00T) reaches 85.0% versus 75.0% for GR00T N1.7, gains of 16.7 and 10.0 points over the bases, as Fig. 5 shows. The gains concentrate on table clearing, obstacle avoidance, and scene tidying, three long-horizon tasks in which several stages must succeed in sequence: ζ_0 (Ψ_0) succeeds in 19 of their 30 trials versus 11 for Ψ_0, and ζ_0 (GR00T) in 23 versus 19 for GR00T N1.7. For ζ_0 (Ψ_0), scene tidying rises from 20 to 60% and table clearing from 50 to 80%; for ζ_0 (GR00T), table clearing rises from 70 to 90%; on the other three tasks, every policy succeeds in 8 to 10 of 10 trials. We observed that small deviations in the relative placement of the body and objects tended to accumulate from stage to stage in baseline trials until a later stage failed, whereas ζ_0 executed more stably across stages. ζ_0 (GR00T) matches or exceeds the other four policies on every task; Fig. 1(a) shows one of its scene-tidying episodes.

Model size and latency. The deployed ζ_0 (Ψ_0) has 2.59B parameters, of which 2.13B form the frozen backbone and 461M the task modules. On one RTX 4090, a policy call takes 212 ms on average and 221 ms at the 95th percentile, using 6.3 GiB, excluding sensing, networking, and control; measurement conditions are in Appendix A.5.

5. Conclusion

ζ_0 is a framework that turns vision-language-action policies into World Action Models by learning joint scene-and-body foresight together with them. A shared latent representation captures future visual features and measured body trajectories, and dual-response transfer supervises its causal prediction in action space through a recorded-future teacher and the executing policy. Built on Ψ_0 and GR00T N1.7, ζ_0 raises mean success over both bases by 18.8 and 9.0 points on seven HumanoidArena tasks and by 16.7 and 10.0 points on six real-world tasks with 10 trials each. Removing future conditioning causes the largest drop on Ψ_0, whereas removing body futures causes the largest drop on GR00T N1.7; dual-response transfer improves both bases. Foreseeing the body as well as the scene, and judging that foresight by the actions it induces, lets a policy anticipate the consequences of its own whole-body commands.

Future work. Pretraining the shared scene-and-body representation on large, diverse humanoid and human-motion data offers a direct path to broaden foresight across tasks and embodiments; combining it with explicit planning or language-level reasoning over longer task structure is a natural next step. Dual-response transfer could also supervise other predictive signals, such as future touch or force.

References

[1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, et al. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[2] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[4] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, et al. π_0: A vision-language-action flow model for general robot control. In *Robotics: Science and Systems*, 2025.

[5] Kevin Black, Manuel Y. Galliker, and Sergey Levine. Real-time execution of action chunking flow policies. In *Advances in Neural Information Processing Systems*, 2025.

[6] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, 2023.

[7] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, et al. RT-1: Robotics transformer for real-world control at scale. In *Robotics: Science and Systems*, 2023.

[8] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, et al. Genie: Generative interactive environments. In *International Conference on Machine Learning*, 2024.

[9] Zhanxiang Cao, Liyun Yan, Yang Zhang, Sirui Chen, Jianming Ma, Tianyue Zhan, et al. HiWET: Hierarchical world-frame end-effector tracking for long-horizon humanoid loco-manipulation. *arXiv preprint arXiv:2602.06341*, 2026.

[10] Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In *Conference on Robot Learning*, 2019.

[11] Yizhi Chen, Zhanxiang Cao, Xinyi Peng, Yixiao Zheng, Xiaxi Si, Yiheng Li, et al. GeoAlign: Beyond semantics with state-guided spatial alignment in VLA models. *arXiv preprint arXiv:2606.03240*, 2026.

[12] Yi Chen, Yuying Ge, Hui Zhou, Mingyu Ding, Yixiao Ge, and Xihui Liu. DIAL: Decoupling intent and action via latent world modeling for end-to-end VLA. *arXiv preprint arXiv:2603.29844*, 2026.

[13] Yizhi Chen, Zheng Zhang, Zhanxiang Cao, Yihe Chen, Shengcheng Fu, Liyun Yan, et al. HierKick: Hierarchical reinforcement learning for vision-guided soccer robot control. *arXiv preprint arXiv:2603.00948*, 2026.

[14] Xuxin Cheng, Yandong Ji, Junming Chen, Ruihan Yang, Ge Yang, and Xiaolong Wang. Expressive whole-body control for humanoid robots. In *Robotics: Science and Systems*, 2024.

[15] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin CM Burchfiel, et al. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems*, 2023.

[16] Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, et al. Learning universal policies via text-guided video generation. In *Advances in Neural Information Processing Systems*, 2023.

[17] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. HumanPlus: Humanoid shadowing and imitation from humans. In *Conference on Robot Learning*, 2024.

[18] GigaWorld Team. GigaWorld-Policy-0.5: A faster and stronger WAM empowered by AutoResearch. *arXiv preprint arXiv:2607.13960*, 2026.

[19] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020.

[20] Qiwen Gu, Jifan Li, Bingjie Gao, Rui Chen, Jing Tang, Xiangxiang Chu, et al. The right future for action: Learning action-relevant predictive states in world action models. *arXiv preprint arXiv:2609.23369*, 2026.

[21] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[22] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

[23] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 2025.

[24] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, et al. OmniH2O: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *Conference on Robot Learning*, 2024.

[25] Tairan He, Jiawei Gao, Wenli Xiao, Yuanhang Zhang, Zi Wang, Jiashun Wang, et al. ASAP: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. In *Robotics: Science and Systems*, 2025.

[26] Tairan He, Wenli Xiao, Toru Lin, Zhengyi Luo, Zhenjia Xu, Zhenyu Jiang, et al. HOVER: Versatile neural whole-body controller for humanoid robots. In *IEEE International Conference on Robotics and Automation*, 2025.

[27] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[28] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, et al. Video prediction policy: A generalist robot policy with predictive visual representations. In *International Conference on Machine Learning*, 2025.

[29] Bingqi Huang, Bingchuan Wei, Xuan Wang, Yingkai Cai, and Zhaokui Wang. Cross-view action consistency for camera-robust vision-language-action policies. *arXiv preprint arXiv:2608.06965*, 2026.

[30] Mazeyu Ji, Xuanbin Peng, Fangchen Liu, Jialong Li, Ge Yang, Xuxin Cheng, et al. ExBody2: Advanced expressive humanoid whole-body control. *arXiv preprint arXiv:2412.13196*, 2024.

[31] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, et al. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning*, 2024.

[32] Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. RMA: Rapid motor adaptation for legged robots. In *Robotics: Science and Systems*, 2021.

[33] Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

[34] Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. In *Robotics: Science and Systems*, 2025.

[35] Yanjun Li, Jiarui Ye, Weiyu Guo, Zixiao He, Kangjie Ye, Dihan Wang, et al. Language-Forcing: Advancing humanoid loco-manipulation via semantic alignment of whole-body world action models. *arXiv preprint arXiv:2610.23894*, 2026.

[36] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

[37] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

[38] Yang Liu, Jiacheng Liu, Hengtao Li, Yuheng Huang, Zijian Yang, Ji Dai, et al. HuGe: Scaling humanoid generalist vision-language-action models with egocentric experience. *arXiv preprint arXiv:2610.34186*, 2026.

[39] Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, et al. SONIC: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 2026.

[40] NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

[41] NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, et al. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[42] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, et al. Octo: An open-source generalist robot policy. In *Robotics: Science and Systems*, 2024.

[43] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023.

[44] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. In *Conference on Robot Learning*, 2025.

[45] Carmelo Sferrazza, Dun-Ming Huang, Xingyu Lin, Youngwoon Lee, and Pieter Abbeel. HumanoidBench: Simulated humanoid benchmark for whole-body locomotion and manipulation. In *Robotics: Science and Systems*, 2024.

[46] Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, et al. Predictive inverse dynamics models are scalable learners for robotic manipulation. In *International Conference on Learning Representations*, 2025.

[47] Taowen Wang, Zikang Xie, Bin Yang, Yunheng Wang, Zizhao Yuan, Yuetong Fang, et al. HumanoidArena: Benchmarking egocentric hierarchical whole-body learning. *arXiv preprint arXiv:2606.17833*, 2026.

[48] Yunheng Wang, Jiaxi Zhang, Wei Lu, Yuetong Fang, Taowen Wang, Yixiao Feng, et al. Kimodo-Policy: Learning kinematic action models for humanoid loco-manipulation. *arXiv preprint arXiv:2611.11554*, 2026.

[49] Songlin Wei, Hongyi Jing, Boqian Li, Zhenyu Zhao, Jiageng Mao, Zhenhao Ni, et al. Ψ_0: An open foundation model towards universal humanoid loco-manipulation. *arXiv preprint arXiv:2603.12263*, 2026.

[50] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, et al. Unleashing large-scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations*, 2024.

[51] Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-WAM: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.

[52] Junpeng Yue, Boyuan Li, Yuxuan Wang, Zepeng Wang, Yuhui Fu, Feiyang Xie, et al. Being-M0.7: A latent world-action model for humanoid robots. *arXiv preprint arXiv:2610.11283*, 2026.

[53] Yanjie Ze, Zixuan Chen, João Pedro Araújo, Zi-ang Cao, Xue Bin Peng, Jiajun Wu, et al. TWIST: Teleoperated whole-body imitation system. In *Conference on Robot Learning*, 2025.

[54] Yanjie Ze, Zixuan Chen, Wenhao Wang, Tianyi Chen, Xialin He, Ying Yuan, et al. Generalizable humanoid manipulation with 3D diffusion policies. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2025.

[55] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems*, 2023.

[56] Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, et al. FLARE: Robot learning with implicit world modeling. In *Conference on Robot Learning*, 2025.

[57] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *International Conference on Machine Learning*, 2025.

[58] Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. In *Robotics: Science and Systems*, 2025.

ζ_0: A Framework for World Action Models with Joint Scene and Body Foresight in Humanoid Control

Appendix

A. Implementation Details

This appendix gives the initialization, losses and parameter updates, architecture and compute, action execution, and simulation-baseline settings of ζ_0 on both bases, followed by real-robot details, per-task ablation results, and body-forecast evaluation details.

A.1. Initialization and vision-language backbone

In ζ_0 (Ψ_0), all task modules, in simulation and on the real robot, are randomly initialized. Its simulation and real-robot versions use the Qwen3-VL-2B-Instruct vision-language model released with Ψ_0, frozen throughout. Each camera image keeps its aspect ratio, is resized with bilinear antialiasing to a short side of at most 240 pixels without upscaling, and is padded symmetrically with RGB value 127 to a multiple of 32; pixels are rescaled by $1/255$ and normalized with mean and standard deviation 0.5, and the backbone's own resizing is disabled. Each image yields at most 160 merged visual tokens, which the backbone's context sequence keeps unchanged. The current-visual branch and the future visual targets use post-merger features averaged over a 4×4 grid with area weights over the valid image region, giving 16 tokens of 2048 dimensions per image; pure padding receives zero weight (Fig. S1a).

Table S1. **Shared seven-task training settings of ζ_0 (Ψ_0).** All eight ζ_0 (Ψ_0) settings use the training recipe of the full model, with the intervention specified in Table 2.

Setting	Value
Tasks / demonstrations per task	7 / 100
Shared model / total demonstrations	1 / 700
Task sampling	Uniform $(1/7)$
Task-module initialization	Random
Vision-language backbone	Frozen
Training updates	40,000
Training seed	292285
GPUs / global batch	2 / 512
Optimizer / Adam betas	AdamW / $(0.95, 0.999)$
World / action learning rates	3×10^{-5} / 10^{-4}
Weight decay / gradient norm clip	10^{-6} / 1
Learning-rate warmup updates	1,000
Learning-rate schedule	Cosine, horizon extended from 20k to 40k at update 20k
Target-encoder EMA decay	0.99
Max committed prefix (training)	15 steps

In simulation, ζ_0 (GR00T) is built inside the native GR00T N1.7 architecture. It keeps the frozen Qwen-based vision-language backbone of the GR00T N1.7 checkpoint (16 text layers, 2048-dimensional context) and its flow-matching DiT action head (32 layers, preceded by four vision-language self-attention layers). The DiT, vision-language normalization and self-attention, and position embeddings load from the checkpoint; a new embodiment-specific state encoder, action encoder, and action decoder handle the 64-dimensional state and 40-dimensional semantic actions. The teacher T is initialized from this head, and the executing policy G starts as an independent copy with separate parameters; the future encoders, readout, conditioner, predictor, and future projections are newly initialized. The three latent groups of eight tokens are projected to the native context width and appended to the vision-language context of the DiT cross-attention, with one masked copy for the image-attention blocks and one for the text-attention

Table S2. **Representation, loss, and execution settings of ζ_0 (Ψ_0) in simulation.** Response coefficients refer to the full model; each single-reader control has unit weight. Scenario seeds vary evaluation conditions, not training replicas. Differences for ζ_0 (GR00T) are specified in Sec. A.1.

Setting	Value
Future-group endpoints (s)	0.2, 0.4, 0.6
Tokens / group; channels / token	8; 256
Measured body targets	29 joints
Native state / action dimensions	64 / 40
Teacher / executing response weight	0.5 / 0.5
Outer response / body weight	1 / 1
Scene / temporal loss weight	1 / 0.25
Forecast-readout / regularizer weight	0.25 / 10^{-3}
Prior / world latent MSE weight	1 / 1
Model action grid	50 Hz
Generated / planned action steps	50 / 30
Executed / committed action steps	24 / 6
Action flow integration steps	10
Scenario seeds	0, 1, 2
Episodes per seed / per task	20 / 60
Episodes per setting	420
Task aggregation	Equal-weight mean
Low-level controller	SONIC
Episode limit: double desk / football	2000 / 2000
Episode limit: box / boxing	1450 / 900
Episode limit: open door / sit sofa	1800 / 2000
Episode limit: navigation	1800

Table S3. **Gradient paths during joint learning.** A checkmark indicates a parameter update from that loss. H/F includes causal conditioning adapters. Fixed readers preserve latent-input derivatives. EMA targets and the frozen backbone receive no gradient.

Loss	Online E	Readout D	Teacher T	Conditioner / predictor H/F	Policy G
Recorded-future reconstruction	✓	✓	—	—	—
Latent regularization	✓	—	—	—	—
Teacher action loss	✓	—	✓	—	—
Latent forecasting	—	—	—	✓	—
Forecast readout	—	fixed	—	✓	—
Dual-response transfer	—	—	fixed	✓	fixed
Executing-policy action loss	—	—	—	✓	✓

blocks; the upstream attention code is unchanged. The head generates 40-step chunks with four Euler steps. The noise fraction τ in Sec. 3.4 maps to the native flow time $1 - \tau$, ordinary action learning uses the native Beta time distribution, and the committed prefix stays clean at every step. Image preprocessing, visual and body targets, loss weights, learning rates, warmup, EMA decay, training seed, and the dual-response objective match ζ_0 (Ψ_0) (Tables S1 and S2), with global batch 512 on two GPUs. The seven ablations apply the same interventions as for ζ_0 (Ψ_0), with matched data, initialization, and 40,000-update budget.

A.2. Losses, masks, and parameter updates

Reconstruction errors are averaged over valid elements per sample and combined with their loss weights; body errors are additionally weighted per joint. Temporal scene differences require valid visual targets in consecutive groups. Body differences additionally require valid corresponding query slots in the two groups. World prediction uses a cumulative action-coverage mask: a missing action segment invalidates the affected group and subsequent world groups. This eligibility rule applies to world losses; the example remains available for ordinary policy learning when its own targets are valid.

The latent regularizer is

$$\mathcal{R}(z) = \frac{1}{|\mathcal{V}|} \sum_{g \in \mathcal{V}} \left([r_g - 4]_+^2 + [0.1 - d_g]_+^2 \right), \qquad \text{(S1)}$$

where $\mathcal{V}$ is the set of valid groups, r_g is the root mean square across that group's latent entries, and d_g is the mean channel-wise standard deviation across its eight tokens. This penalizes excessive scale and identical tokens.

Table S3 summarizes parameter updates during joint learning. The online encoder receives reconstruction and teacher-action gradients. EMA targets are stop-gradient labels. Forecast readout and response losses hold their readers fixed while retaining input derivatives. The executing-policy flow-matching loss updates both G and its causal prediction path.

A.3. Joint training algorithm

Algorithm S1 summarizes joint training, with parameter gradients following Table S3. Prior and world forecasting alternate by optimizer update; the executing policy uses the causal prior prediction in both modes. The EMA encoder advances once at the start of each update, before target encoding, including when gradients are accumulated over multiple microbatches.

A.4. Architecture and compute

Table S4 lists the transformer modules and parameter counts of ζ_0 (Ψ_0). Within ζ_0 (Ψ_0), widths and depths are identical in simulation and on the real robot; counts differ only through the state and action adapters. During training its task modules hold 582M parameters, of which 563M are trainable and 19M form the EMA target encoder. Each ζ_0 (Ψ_0) simulation setting trains on two 80 GB GPUs; 40,000 updates take about 37 h for the full model and 36–43 h across all eight settings. The full ζ_0 (GR00T) model trains on two 141 GB NVIDIA H200 GPUs and takes about 50 h for 40,000 updates. Both times exclude feature caching.

Table S4. **Architecture and parameter counts of ζ_0 (Ψ_0).** Transformer widths and depths are shared by the simulation and real-robot models. Parameter counts exclude normalization buffers.

Module	Width	Heads	FFN	Layers
Conditioner H	1024	16	4096	16
Predictor F	768	12	3072	8
Future encoder E	512	8	2048	4
Teacher T, policy G (each)	768	12	3072	8

Parameters	Simulation	Real robot
Frozen backbone	2,131,333,120	2,131,333,120
Task modules, training	581,809,687	582,013,279
trainable	562,679,063	562,882,655
Task modules, deployed	461,064,637	461,191,137
Deployed total	2,592,397,757	2,592,524,257

Algorithm S1 Joint training of ζ_0

Require: Demonstrations (x_t, p_t, a_t, y_t) with validity masks; modules $E_\phi, D, H, F_\theta, T_\psi, G_\omega$; EMA decay $\rho = 0.99$.

Initialize modules as in Sec. A.1; set $\bar{\phi} \leftarrow \phi$ and freeze the vision-language backbone.
for optimizer update $k = 0, 1, \ldots$ **do**
  Set $m \leftarrow$ prior if k is even, and $m \leftarrow$ world otherwise.
  Advance the EMA once: $\bar{\phi} \leftarrow \rho\bar{\phi} + (1-\rho)\phi$.
  Sample a batch and construct the valid future, action-suffix, and mode-specific masks (Sec. A.2).
  Encode $z_t \leftarrow E_\phi(x_t^{\mathrm{vb}}, y_t)$ and $z_t^\star \leftarrow \mathrm{sg}(E_{\bar{\phi}}(x_t^{\mathrm{vb}}, y_t))$.
  Compute $h_t \leftarrow H(x_t)$ and the causal prediction $\widehat{z}_t^{\mathrm{prior}} \leftarrow F_\theta(h_t, p_t; \mathrm{prior})$.
  if $m =$ world **then**
    Compute $\widehat{z}_t^{\mathrm{world}} \leftarrow F_\theta(h_t, a_t; \mathrm{world})$.
  end if
  Compute recorded-future reconstruction $\mathcal{L}_{\mathrm{rec}}(z_t; D)$ and regularization $\mathcal{R}(z_t)$.
  Compute ordinary flow-matching losses: $\mathcal{L}_T$ conditions on z_t, and $\mathcal{L}_G$ on $\widehat{z}_t^{\mathrm{prior}}$.
  Compute $\mathcal{L}_{\mathrm{pred}}^m = \lambda_m \|\widehat{z}_t^m - z_t^\star\|_{M_m}^2 + \lambda_D \mathcal{L}_{\mathrm{rec}}(\widehat{z}_t^m; D_{\mathrm{fix}})$.
  Set $\mathcal{L}_{\mathrm{resp}} \leftarrow 0$.
  if $m =$ prior **then**
    Draw one shared flow sample X_r for all three responses and detach its inputs.
    Form the reference $v_{\mathrm{ref}} \leftarrow \mathrm{sg}(T_\psi(X_r, z_t^\star))$.
    Compute $\mathcal{L}_{\mathrm{resp}} \leftarrow \frac{1}{2}\|T_{\mathrm{fix}}(X_r, \widehat{z}_t^{\mathrm{prior}}) - v_{\mathrm{ref}}\|_{M_r}^2 + \frac{1}{2}\|G_{\mathrm{fix}}(X_r, \widehat{z}_t^{\mathrm{prior}}) - v_{\mathrm{ref}}\|_{M_r}^2$.
  end if
  Sum $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{reg}}\mathcal{R} + \mathcal{L}_T + \mathcal{L}_G + \mathcal{L}_{\mathrm{pred}}^m + \lambda_r \mathcal{L}_{\mathrm{resp}}$.
  Backpropagate with the gradient paths in Table S3, clip gradients, and take an AdamW step.
end for
return the frozen backbone and H, F_θ, G_ω for causal deployment (Sec. A.5).

A.5. Action sampling and execution

In the action-flow convention of Sec. 3.4, clean data is at time zero and Gaussian noise at time one, with velocity target equal to noise minus the normalized action. ζ_0 (Ψ_0) integrates backward from one to zero in ten steps; ζ_0 (GR00T) in simulation uses four Euler steps in its native time convention, as described above. A committed prefix remains clean throughout sampling and is excluded from the action loss. The predictor estimates the future latent once per chunk; it has no flow-time input (Fig. S1b).

In HumanoidArena, the policy plans 30 action steps, executes 24, and keeps the last six steps of the preceding chunk as the committed prefix.

The real-robot latency of ζ_0 (Ψ_0) in Sec. 4.5 was measured after the trials on one RTX 4090 with batch 1, three 640×480 views, a BF16 backbone, and ten flow steps. One call generates a 50-step chunk, from which execution takes a 30-step window; it takes 212 ms on average (p95 221 ms) with no prefix and 216 ms with a 15-step committed prefix (deployment uses six), using 6.3 GiB, which fits well within the 480 ms (24-step) execution window under receding-horizon execution.

A.6. Simulation baselines

ACT, $\pi_{0.5}$, and Ψ_0 are trained jointly on the same 700 demonstrations for 40,000 updates with global batch 512 and are evaluated at their final checkpoints under the common protocol. $\pi_{0.5}$ and Ψ_0 start from their pretrained checkpoints, and $\pi_{0.5}$ is fine-tuned with all parameters. Table S5 lists their optimizer settings; the GR00T N1.7 baseline uses the same training settings as ζ_0 (GR00T) (Sec. A.1).

B. Real-Robot Details

All five policies use a Unitree G1 with two dexterous hands, current snapshots from three RGB cameras, and a 41-dimensional measured state containing 29 body joints and 12 hand joints. They output a 76-dimensional action of 64 whole-body motion commands for SONIC and 12 hand commands. Execution follows the simulation protocol: plan 30 steps, execute 24, and keep the last six as the committed prefix of the next chunk. Policies read only observations available at decision time; recorded committed actions serve only as training labels. Each policy is trained

Table S5. **Optimizer settings of the simulation baselines.** The learning rate is the initial value, or the peak value after warmup.

Setting	ACT	$\pi_{0.5}$	Ψ_0
Global batch	512	512	512
Learning rate	10^{-5}	2.5×10^{-5}	10^{-4}
Optimizer	AdamW	AdamW	AdamW
Adam betas	$(0.9, 0.999)$	$(0.9, 0.95)$	$(0.95, 0.999)$
Weight decay	10^{-4}	0.01	10^{-6}
Warmup updates	—	1,000	1,000

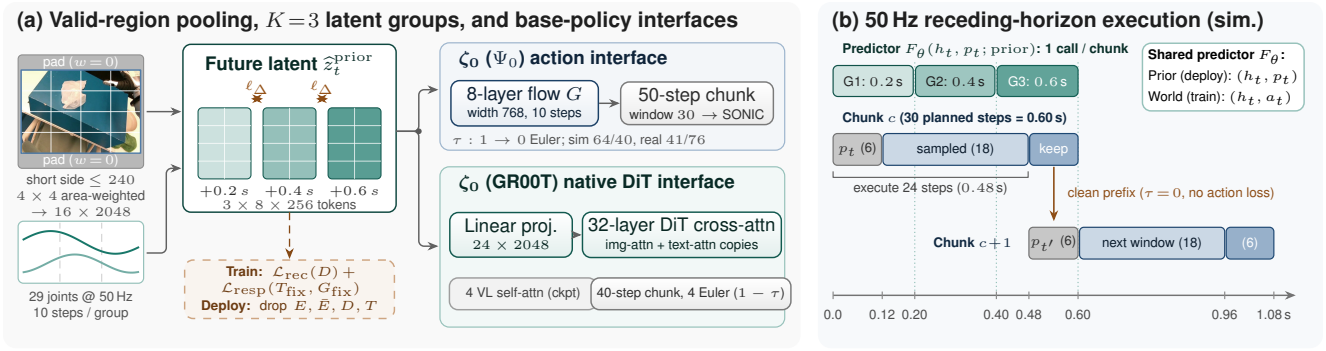


Figure S1. **Representation pooling, base-policy interfaces, and receding-horizon execution of ζ_0. (a)** Aspect-preserved images with zero-weighted padding are pooled on a 4×4 valid-region grid and paired with 50 Hz 29-joint trajectories across $K = 3$ groups to form $\hat{z}_t^{\text{prior}} \in \mathbb{R}^{3\times8\times256}$, which conditions either the 8-layer flow policy G in ζ_0 (Ψ_0) or the 32-layer DiT cross-attention in ζ_0 (GR00T). **(b)** At 50 Hz, each 30-step (0.60 s) planning window matches the $+0.2/0.4/0.6$ s forecast groups: the controller executes 24 steps (0.48 s) and retains the final six (0.12 s) as the clean $\tau = 0$ committed prefix of the next chunk, while F_θ runs once per chunk.

independently for each task on the same demonstrations, using a global batch of 512 and 20,000 updates, and is evaluated at its final checkpoint. Architecture and base-policy initialization follow the respective methods. Both ζ_0 models use the full dual-response objective (teacher and executing-policy response weights 0.5 each), body auxiliary weight 1, and temporal weight 0.25; future-body supervision uses 29 measured body joints. Timing gaps in the recordings split episodes into continuous segments, and future supervision does not cross segment boundaries. In experiments involving people, informed consent was obtained for participation and for the use of their images in the paper. Identifiable faces in the experimental images are obscured. Tables S6 and S8 list the demonstrations, instructions, and common success criteria.

Instructions are reproduced verbatim from the recorded data; the doll task keeps 100 of 101 demonstrations. Each policy is evaluated for 10 trials per task with randomized initial robot and object positions, and a failure or timeout is unsuccessful, giving 60 trials per policy and 300 in total. Table S7 lists the per-task success counts behind the success rates in Fig. 5. Figure S2 shows synchronized third-person and head-camera views across an autonomous scene-tidying rollout, the physical span of one 0.6 s forecast horizon during grasping, and success grouped by stage complexity: while all five policies reach 80.0%–93.3% across the three single-stage tasks (30 trials each), ζ_0 raises success on the three multi-stage tasks (Clear table, Avoid obstacle, Tidy scene; 30 trials each) from 36.7% to 63.3% on Ψ_0 (+26.7 pp) and from 63.3% to 76.7% on GR00T N1.7 (+13.3 pp).

Table S6. **Real-world task data.** All five policies train on the same demonstrations of each task.

Task	Demos	Language instruction
Hair drying	81	Blow-dry the person's hair using a hair dryer.
Push chair	101	Place the chair in front of the table
Clear table	86	Tidy up the table and put the trash into the bin.
Avoid obstacle	100	Place the tape on the table.
Place doll	100	place doll to plate
Tidy scene	200	Tidy up the scene.

C. Per-Task Ablation Results

Figure S3 shows how each ablation changes success on every task relative to the full model on the same base, using the abbreviations of Table 2. Table S9 lists the corresponding per-task success rates; differences Δ in Table 2 are computed from unrounded episode counts. Note that w/o FC retains the full parameter capacity of ζ_0 while zeroing G's future input, yet drops close to each base policy (47.1% vs. 44.8% on Ψ_0, and 73.3% vs. 70.5% on GR00T N1.7), confirming that the gains come from predicted foresight rather than added parameters. On ζ_0 (Ψ_0), removing future conditioning in G (w/o FC) lowers success on six of seven tasks, by up to 23.3 points on pick and place box. On ζ_0 (GR00T),

Table S7. **Real-world success counts.** Each entry is the number of successful trials out of 10; Mean is the equal-weight six-task success rate (%).

Policy	Hair drying	Push chair	Clear table	Avoid obst.	Place doll	Tidy scene	Mean
$\pi_{0.5}$	8	8	4	1	8	1	50.00
Ψ_0	9	8	5	4	9	2	61.67
ζ_0 (Ψ_0)	10	9	8	5	9	6	**78.33**
GR00T N1.7	9	9	7	6	8	6	75.00
ζ_0 (GR00T)	10	9	9	7	9	7	**85.00**

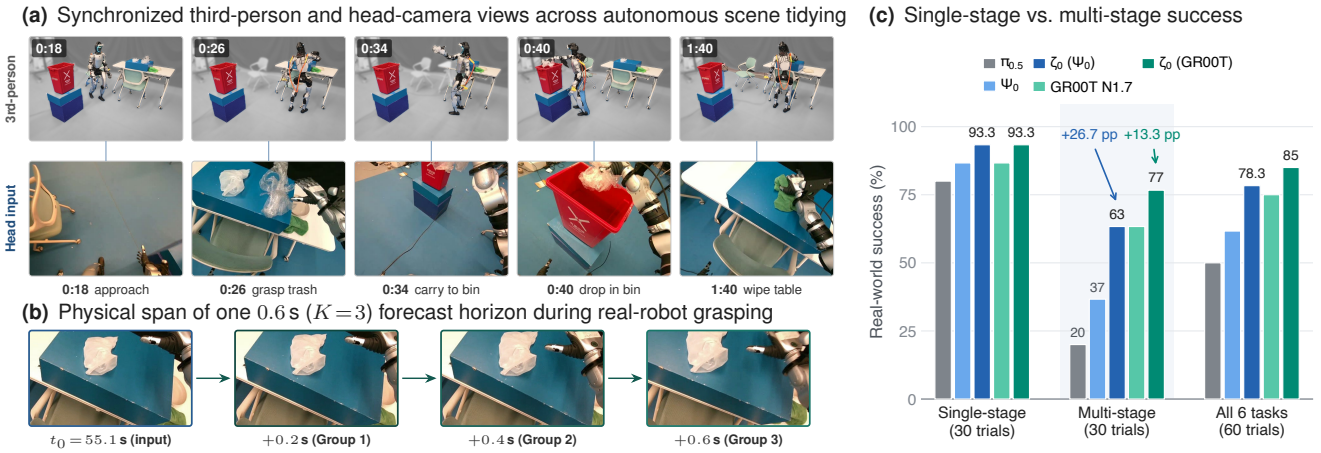


Figure S2. **Synchronized real-robot views, 0.6 s horizon span, and multi-stage task breakdown on the Unitree G1. (a)** Third-person frames (top) paired with synchronized model-input head-camera views (bottom) across an autonomous ζ_0 (GR00T) scene-tidying episode; legs and feet remain outside the head camera's field of view throughout locomotion and manipulation. **(b)** Four consecutive real-robot head-camera frames spaced by 0.2 s, illustrating the interaction span of one $K = 3$ ($+0.2/0.4/0.6$ s) forecast window. **(c)** Success rates grouped into single-stage tasks (Hair drying, Push chair, Place doll; 30 trials per policy) and multi-stage tasks (Clear table, Avoid obstacle, Tidy scene; 30 trials per policy), where ζ_0 raises Ψ_0 by $+26.7$ pp and GR00T N1.7 by $+13.3$ pp.

removing dual-response transfer, the teacher response term, or future conditioning (w/o DR, TR, FC) each lowers success on six of seven tasks; the largest single drop is 20.0 points on navigation without future conditioning. Other per-task changes are smaller and mixed in sign.

Figure S4 summarizes three patterns across benchmarks and ablations. First, across all 13 simulation and real-world tasks on both bases (26 base-to-ζ_0 pairs), ζ_0 improves or matches its base policy in 24 pairs (Fig. S4a). Second, on ζ_0 (GR00T), tasks requiring contact and whole-body coordination—football, box, double desk, and boxing—lose 10.0–13.3 pp when measured body futures are removed (w/o BF), whereas long-horizon navigation loses 20.0 pp when future conditioning is removed from G (w/o FC; Fig. S4b). Third, teacher-only supervision favors precision box placement while executing-policy-only supervision favors reactive boxing on both bases, and combining both readers in dual-response transfer yields the highest seven-task mean on ζ_0 (Ψ_0) (63.6%) and ζ_0 (GR00T) (79.5%; Fig. S4c), while the synchronized benchmark views in Fig. S4d illustrate how egocentric inputs omit lower-body support across all seven tasks.

Table S8. **Real-world task success criteria.** Each of the five policies is evaluated for 10 trials per task.

Task	A trial succeeds if the robot…
Hair drying	grasps the dryer once a person hands it over, dries the person's hair for more than 5 s, and puts the dryer down.
Push chair	walks to the chair, grasps it, and pushes it under the table in an aligned position.
Clear table	drops the two objects on the table into the bin one after the other.
Avoid obstacle	picks up the tape, walks around the obstacle, and places the tape on the target table.
Place doll	walks to the table, grasps the doll, and places it on the plate.
Tidy scene	moves the chair, discards two pieces of trash, wipes the table, and places the cloth on the plate.

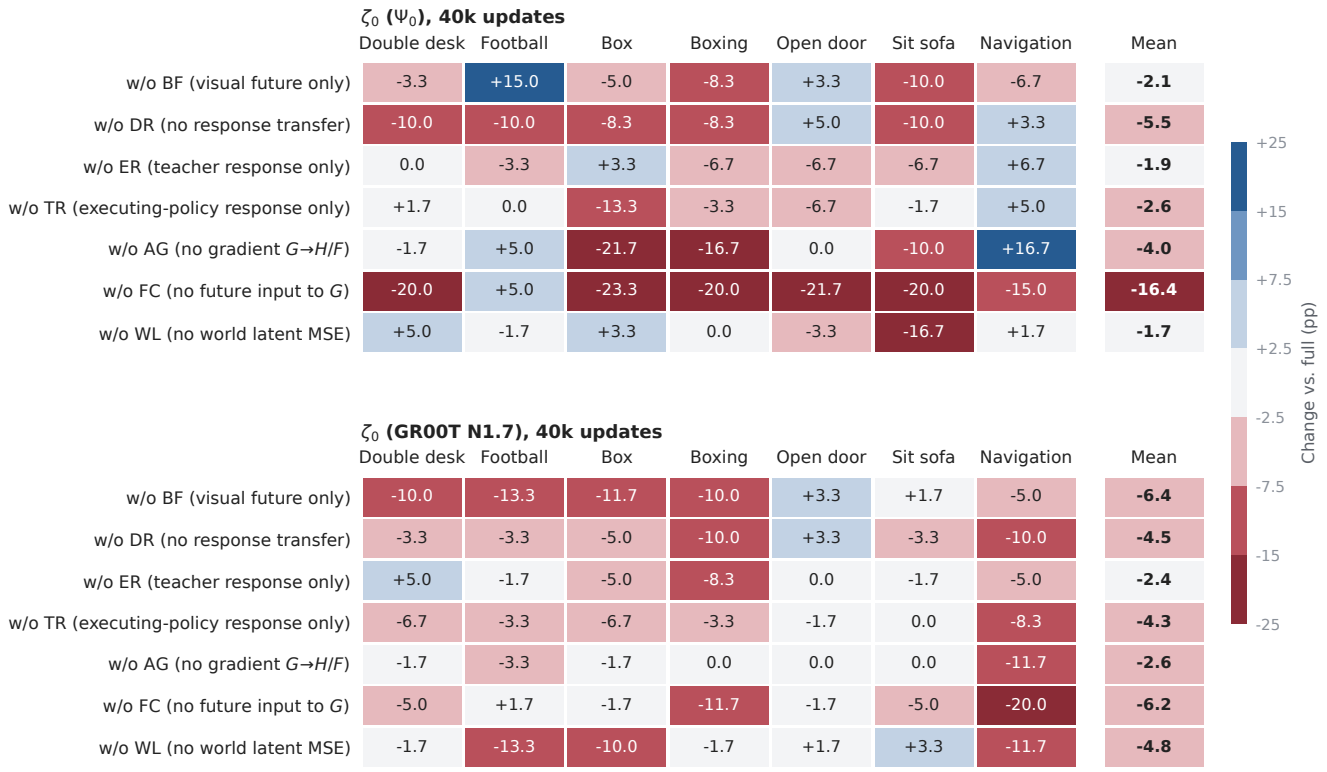


Figure S3. **Per-task change of each ablation relative to the full model on the same base.** Top: ζ_0 (Ψ_0); bottom: ζ_0 (GR00T). Entries are success-rate differences in percentage points over 60 episodes per task; the last column is the change in the seven-task mean, as in Table 2. Per-task success of each full model is listed in Table 1. Colors are binned.

Table S9. **Per-task ablation success of ζ_0 on both bases.** Success rate (%) of each shared seven-task model over 60 episodes per task; all settings use 40,000 updates. Abbreviations follow Table 2, and Base denotes the corresponding base policy.

Base	Setting	Double desk	Football	Box	Boxing	Open door	Sit sofa	Navigation	Mean
Ψ_0	Full	21.67	23.33	85.00	61.67	95.00	96.67	61.67	**63.57**
	w/o BF	18.33	38.33	80.00	53.33	98.33	86.67	55.00	61.43
	w/o DR	11.67	13.33	76.67	53.33	100.00	86.67	65.00	58.10
	w/o ER	21.67	20.00	88.33	55.00	88.33	90.00	68.33	61.67
	w/o TR	23.33	23.33	71.67	58.33	88.33	95.00	66.67	60.95
	w/o AG	20.00	28.33	63.33	45.00	95.00	86.67	78.33	59.52
	w/o FC	1.67	28.33	61.67	41.67	73.33	76.67	46.67	47.14
	w/o WL	26.67	21.67	88.33	61.67	91.67	80.00	63.33	61.90
	Base	6.67	26.67	61.67	50.00	65.00	63.33	40.00	44.76
GR00T	Full	45.00	56.67	96.67	88.33	96.67	95.00	78.33	**79.52**
	w/o BF	35.00	43.33	85.00	78.33	100.00	96.67	73.33	73.10
	w/o DR	41.67	53.33	91.67	78.33	100.00	91.67	68.33	75.00
	w/o ER	50.00	55.00	91.67	80.00	96.67	93.33	73.33	77.14
	w/o TR	38.33	53.33	90.00	85.00	95.00	95.00	70.00	75.24
	w/o AG	43.33	53.33	95.00	88.33	96.67	95.00	66.67	76.90
	w/o FC	40.00	58.33	95.00	76.67	95.00	90.00	58.33	73.33
	w/o WL	43.33	43.33	86.67	86.67	98.33	98.33	66.67	74.76
	Base	43.33	38.33	86.67	75.00	98.33	93.33	58.33	70.48

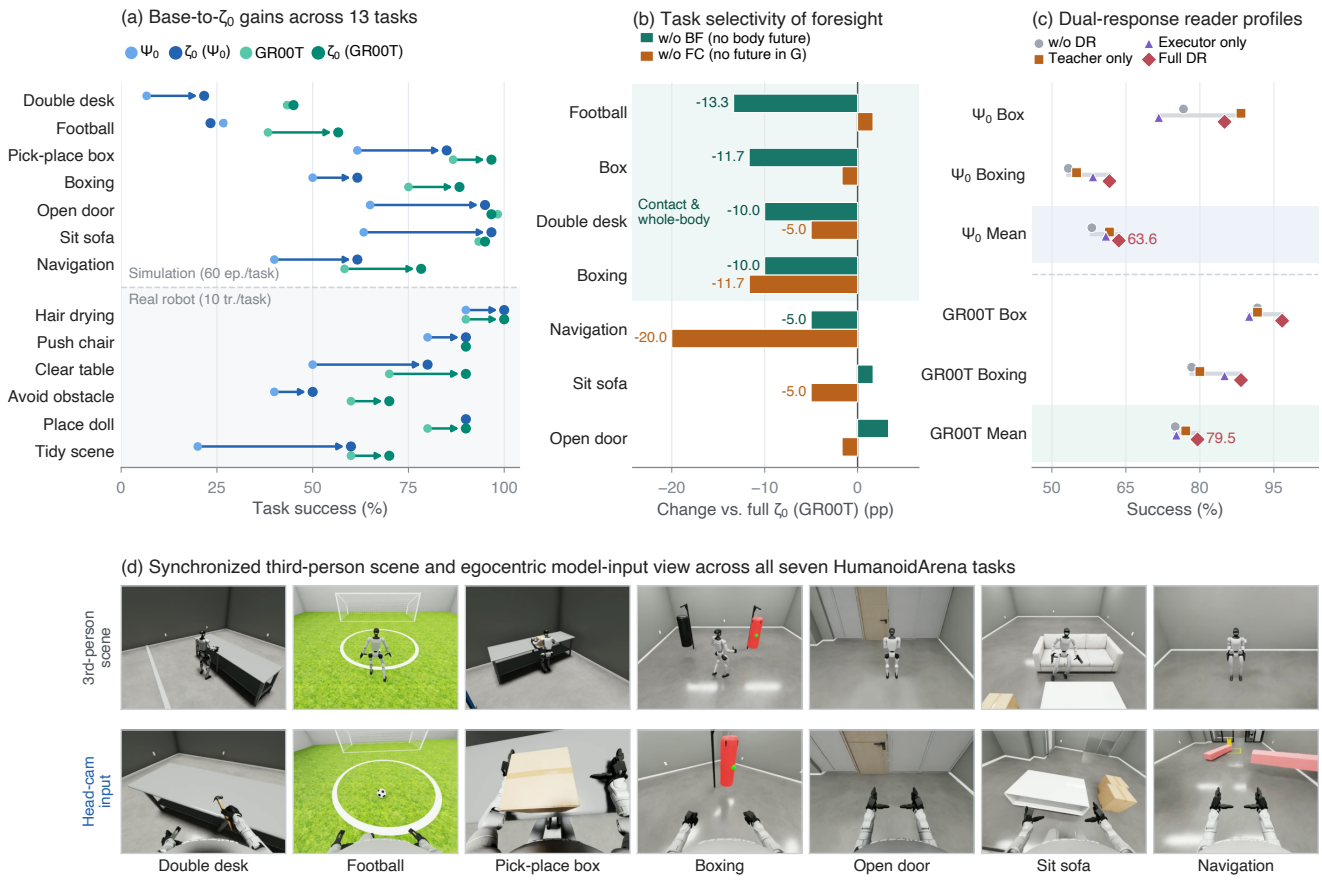


Figure S4. **Cross-benchmark gains, task selectivity, dual-response reader profiles, and HumanoidArena camera views. (a)** Paired base-to-ζ_0 success rates across all seven simulation tasks (60 episodes/task) and six real-world tasks (10 trials/task); ζ_0 improves or matches its base in 24 of 26 task pairs. **(b)** On ζ_0 (GR00T), contact-and-coordination tasks drop most without body futures (w/o BF), whereas navigation drops most without future conditioning in G (w/o FC). **(c)** Across both bases, teacher-only supervision favors precision box manipulation while executor-only supervision favors reactive boxing, and full dual-response transfer achieves the highest seven-task mean. **(d)** Synchronized third-person scene (top) and model-input head camera (bottom) across all seven HumanoidArena tasks; egocentric views show forearms and task objects while the torso, hips, knees, and supporting feet lie outside the field of view.

D. Body-Forecast Evaluation

The body-forecast statistics in Sec. 4.4 use one closed-loop rollout per task of the shared ζ_0 (Ψ_0) checkpoint, fixed before inference, with an anchor every 0.1 s over its first 8 s, giving 526 windows (the boxing rollout ends earlier). Each window forecasts 30 steps at 50 Hz (0.02–0.60 s, or 10 query steps per 0.2 s group) across the 29 measured body joints (14 arm, 12 leg, and 3 waist joints). The scene panels of Fig. 3 are separate windows selected post hoc for display and are not used in these statistics.

Figure S5 and Table S10 report the closed-loop joint trajectories and error breakdowns across all seven tasks and three limb groups. Along continuous rollouts, the causal forecasts anticipate gait peaks in navigation left-knee flexion (10.4° RMSE versus 30.4° for holding the anchor pose), hip swing during football kicking (5.1° versus 17.8°), and right-elbow extension in double-desk manipulation (7.2° versus 12.3°; Fig. S5a–c).

Although the lower limbs are frequently outside the head camera's field of view or occluded by manipulated objects (Fig. S5d), the 12 leg joints incur the largest pose-persistence error across the 526 windows (10.34° pooled and 12.54° at +0.6 s) and gain the largest absolute error reduction from ζ_0 (4.35° pooled and 5.23° at +0.6 s; Fig. S5e). Across all seven tasks, ζ_0 lowers pooled RMSE and wins 83.46% (439/526) of +0.6 s windows (Fig. S5f, Table S10); pick and place box has the smallest drift (5.45°, endpoint win 50.62%), yet still drops to 4.11°.

Limitations. Each simulation setting uses one training seed (420 episodes), so 1–2 pp ablation gaps should be read alongside consistent cross-base trends. Real-world models train per task (10 trials each), body forecasts omit root motion, and the 212 ms latency excludes perception, networking, and SONIC.

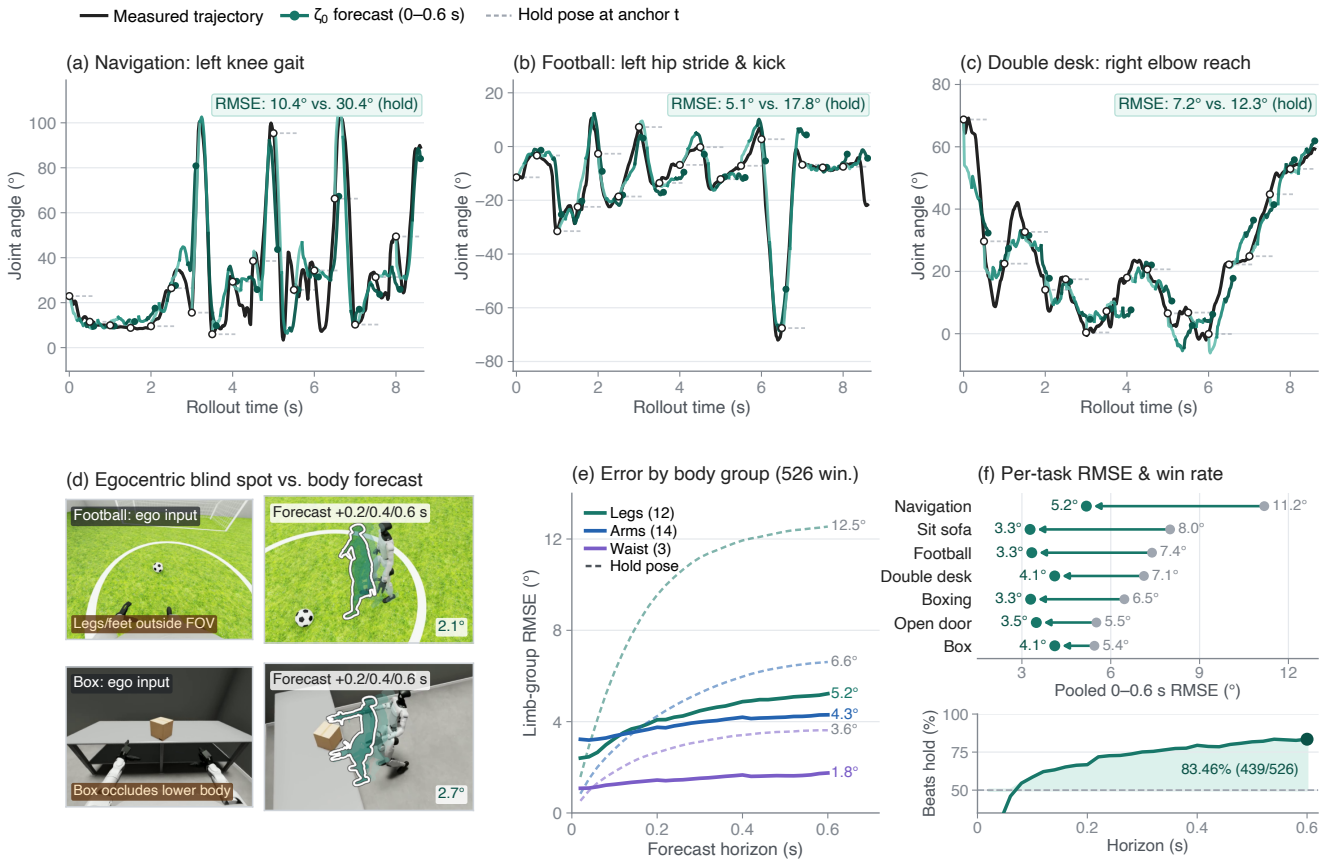


Figure S5. **Closed-loop joint-trajectory forecasts, egocentric blind spots, and per-limb / per-task statistics with ζ_0 (Ψ_0). (a–c)** Measured joint trajectories (black) over the first 8 s of closed-loop rollouts with 0–0.6 s causal forecasts (green, shown every 0.5 s for clarity) versus holding the anchor pose (dashed gray); badges report RMSE over all rollout anchors of that joint. **(d)** Showcase head-camera inputs (left, selected post hoc as in Fig. 3) omit or occlude the lower body during kicking and box manipulation, whereas ζ_0 still anticipates the whole-body motion over $+0.2/0.4/0.6$ s (right; white outline shows the measured $+0.6$ s body). **(e)** Across all 526 windows, the 12 leg joints incur the largest pose-persistence drift (12.5° at $+0.6$ s) and gain the largest absolute error reduction from ζ_0 (5.2°). **(f)** ζ_0 lowers pooled 0–0.6 s RMSE on all seven tasks and beats pose persistence in 83.46% (439/526) of windows at $+0.6$ s.

Table S10. **Per-task and per-limb closed-loop body-forecast RMSE (°) of ζ_0 (Ψ_0) across 526 evaluation windows.** *Hold* keeps the measured 29-joint pose at anchor t. Pooled columns average squared error over all 30 steps (0.02–0.60 s at 50 Hz), and *Win @ $+0.6$ s* is the percentage of windows in which the $+0.6$ s forecast has lower 29-joint RMSE than holding the pose at t.

		Pooled (0–0.6 s)		Endpoint ($+0.6$ s)		Pooled by limb (ζ_0 / Hold)			
Task	Windows	ζ_0	Hold	ζ_0	Hold	Arms (14)	Legs (12)	Waist (3)	Win @ $+0.6$ s (%)
Double desk	81	**4.11**	7.12	**4.66**	9.16	4.62 / 6.17	3.90 / 8.76	1.66 / 2.37	95.06
Football	81	**3.33**	7.39	**3.83**	9.46	3.02 / 5.38	3.96 / 9.70	1.46 / 4.08	95.06
Pick & place box	81	**4.11**	5.45	**4.84**	6.89	4.56 / 4.72	4.02 / 6.69	1.29 / 2.03	50.62
Boxing	40	**3.29**	6.46	**3.89**	7.41	2.92 / 4.40	3.99 / 8.77	1.26 / 2.22	92.50
Open door	81	**3.49**	5.52	**4.23**	6.73	3.64 / 4.14	3.66 / 7.23	1.51 / 2.29	75.31
Sit sofa	81	**3.28**	8.00	**3.80**	9.08	2.62 / 4.01	4.19 / 11.56	1.25 / 3.17	97.53
Navigation	81	**5.18**	11.18	**5.81**	13.90	4.78 / 6.12	6.11 / 15.98	1.81 / 3.56	82.72
All windows	526	**3.92**	7.60	**4.53**	9.36	**3.89** / 5.11	**4.35** / 10.34	**1.49** / 2.96	**83.46**