WHAT 30,000 HOURS OF EGO-CENTRIC VIDEO DOES NOT TEACH

Anonymous authors
Paper under double-blind review

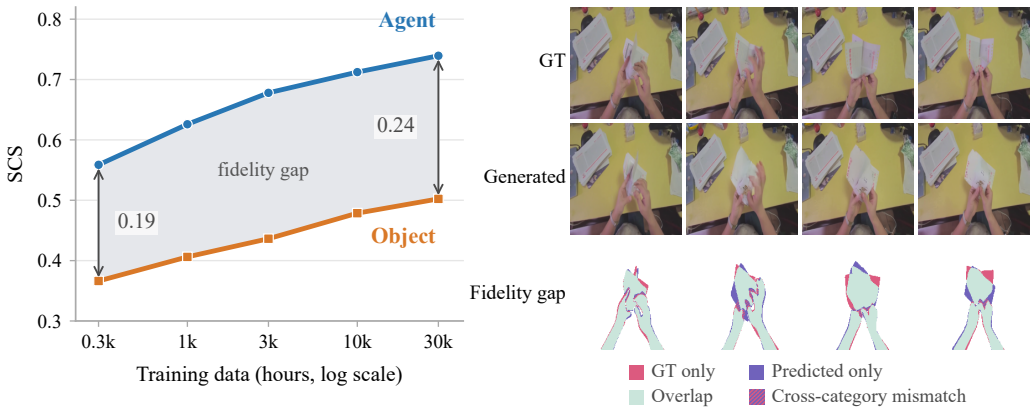

Figure 1: **Limitations of world model scaling.** *Left:* Ego-centric data scaling can achieve high agent modeling fidelity, but the effect of agent's actions on the world lags behind. *Right:* The same failure at the level of a single prediction. Our best model places the hands almost exactly where they belong, while the paper they are folding is rendered in the wrong configuration throughout. See supplementary web page for more video-level results.

ABSTRACT

World models offer a promising alternative to physics-based simulators, yet remain far from practical deployment. We ask how far scaling ego-centric human video takes them, using a dataset of 30,000 hours spanning over 1,000 scene types and 14,000 contributors. Rather than relying on opaque downstream metrics, we directly evaluate agent and object-interaction fidelity on a challenging out-of-distribution benchmark. Increasing training data by 100× improves both, but unevenly: the agent is modeled well, while object fidelity remains far lower and improves slowly. We show that the agent gains need not come from data, and a careful visual conditioning design saturates fidelity with a fraction of it, which lets us measure object fidelity on its own and discover its saturation point. We then introduce a supervision scheme that shifts capacity from scene appearance toward object dynamics, improving object fidelity though a substantial gap remains. Finally, our conclusions transfer to downstream humanoid modeling. Overall, our results suggest that scaling ego-centric data brings agent modeling close to its limit while leaving its effects on the world far behind, and that closing this gap will depend on how models are trained, not only on how much data they see.

1 INTRODUCTION

World models (Schmidhuber, 1990) are being pursued for two distinct roles in embodied AI. As simulators, they promise to replace hand-authored environments for training and evaluating policies at scale, where relying on the physical world alone is prohibitively costly, slow, and unsafe (Guo et al., 2026; Gao et al., 2026). As predictive models for planning, they promise to let an agent

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

evaluate candidate actions at inference time by rolling out their consequences (Ha & Schmidhuber, 2018; Bar et al., 2025). Both roles rest on the same capability: predicting how the objects an agent interacts with respond to its actions. This is precisely where conventional physics-based simulators fall short: contact, friction and deformation resist accurate simulation even when contact parameters are fit to real-world data (Acosta et al., 2022; Blanco-Mulero et al., 2024). World models offer to learn these phenomena directly from data, if they in fact acquire them.

Recent efforts have scaled world models to large and diverse datasets of human–object interaction (Punamiya et al., 2026), in some cases comprising decades of interaction-dense footage (Luo et al., 2026; Dyna Robotics, 2026). Consistent gains with more data and larger models have encouraged a scaling-centric view: that dataset and model size are all that is needed to learn the dynamics of the world. Critics counter that pixel-level losses allocate model capacity by pixel mass and predictability rather than by causal relevance, so that the objects and contacts are precisely what the objective under-weights (LeCun et al., 2022). If correct, this is a failure mode that scaling alone will not efficiently resolve. The argument has motivated alternative, latent-predictive formulations (Assran et al., 2023; Bardes et al., 2024), but the effect it posits has not been measured directly.

Testing that claim requires measuring interaction fidelity itself, across scale. Evidence for the scaling-centric view comes largely from downstream, task-level results: agreement between policy performance in learned and physics-based simulators (Gao et al., 2026), or offline action error when world models are used as policies (Dyna Robotics, 2026). Such evaluations compound errors across perception, agent motion, object dynamics, and policy behavior. We instead directly employ interaction-fidelity metrics (Bagchi et al., 2026; Pallotta et al., 2026), evaluating agent motion and its effect on the world separately as a function of training data, on a new out-of-distribution benchmark.

Specifically, we train Cosmos 3 video world models (NVIDIA, 2026) on nested subsets, from 300 to 30,000 hours, of a large and diverse corpus of ego-centric human manipulation spanning 32k tasks and 116 environment types, collected by 14k participants (see Section 3.2 for details). We repeat this sweep for two model variants and two action-conditioning designs, one of which supplies the agent's body configuration as a projected skeleton (Wang et al., 2025b; Wu & Gao, 2026). All models are evaluated on an out-of-distribution benchmark with different capture hardware, sites, and participants, which we will release publicly.

Our experiments demonstrate that scaling ego-centric data improves both the agent and its action effects, but not equally. As shown in Figure 1, agent fidelity saturates first, while object fidelity lags far behind with the gap between the two widening with more data. Better conditioning substitutes for most of the agent gains: 1k hours of skeleton-conditioned training exceed the agent fidelity that standard conditioning reaches at 30k, and both converge to nearly the same limit (Figure 4, left). Saturating the agent this way isolates action effects and reveals that they converge far below the agent's fidelity and far below the level that is usable for simulation (Figure 4, right).

Recently proposed interaction-focused objectives (Zhang et al., 2026b) do not change this. We introduce object-centric adaptive noise scheduling (Section 5), which shifts supervision away from overall scene appearance and toward the dynamics of manipulated objects, and is the only intervention we test that raises the level object fidelity converges to. The improvement is nonetheless partial: object fidelity remains far below the agent's, and closing the remaining gap will require more than data. Finally, we show that these gains transfer to a downstream humanoid setting (Section 6).

Contributions. We present a controlled study of how interaction fidelity in ego-centric world models behaves as a function of scale. Concretely:

- We train Cosmos 3 world models on up to 30,000 hours of human–object interaction and find that agent fidelity and the fidelity of the agent's effects on the world behave very differently under scale: the agent saturates early and is reached far more cheaply through conditioning than through data, while its effects converge far below it (Section 4).
- We propose object-centric adaptive noise scheduling, which shifts the training signal from scene appearance toward the dynamics of manipulated objects. It is the only intervention we test that raises the level object fidelity converges to, though a substantial gap remains (Section 5).
- We show that these gains transfer to a downstream humanoid setting, where our design contributes more than ego-centric pre-training alone does, but modeling humanoid's interactions with the world remains a bottleneck (Section 6).

2 RELATED WORK

World models for robotics. The idea that an agent can act by consulting an internal model of its environment dates to Craik (1943), and was realized in neural controllers that predict future observations and plan through them (Schmidhuber, 1990; Sutton, 1991) before being revived with modern architectures by Ha & Schmidhuber (2018). In robotics, such models served as forward predictors in MPC, with actions selected by sampling candidate sequences and scoring their outcomes (Finn & Levine, 2017; Ebert et al., 2018), or by optimizing policies within a learned latent space (Hafner et al., 2019; 2023). Their scope was limited to fixed setups with a single embodiment and few objects, by the small scale of available robotics datasets (Dasari et al., 2019) and by early generation objectives (Srivastava et al., 2015; Mathieu et al., 2016) that collapsed multi-modal futures toward a conditional mean and did not scale beyond low resolution and short horizons (Villegas et al., 2019).

Both limitations have since been addressed. Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020) replaced regression with a sampling-based objective, and operating in a compressed latent space (Rombach et al., 2022) with transformer backbones (Peebles & Xie, 2023) made training on large datasets (Chen et al., 2024) tractable, while coordinated collection efforts assembled robot interaction at a scale unavailable to earlier work (O'Neill et al., 2024; Khazatsky et al., 2024). This resulted in a new wave of world models for robotics: IRASim (Zhu et al., 2025) and Ctrl-World (Guo et al., 2026) for tabletop manipulation, NWM (Bar et al., 2025) and EgoWM (Bagchi et al., 2026) for ego-centric navigation and manipulation. What still bounds them is the data: teleoperation remains expensive, and the largest datasets cover only a few hundred scenes (Khazatsky et al., 2024).

Ego-centric human video, in contrast, is abundant, diverse, and requires no robot, and has driven a rapid scale-up of world-model pretraining. DreamDojo (Gao et al., 2026) trains on 44k hours before post-training on small amounts of target robot data, reporting strong action controllability across several humanoid embodiments; Being-H0.7 (Luo et al., 2026) scales to 200k hours in a world-action formulation; and EgoScale (Zheng et al., 2026) uses diverse ego-centric human data to improve dexterous manipulation. Industrial systems report pretraining on larger datasets still (Dyna Robotics, 2026), with the prevailing conclusion that more human video yields better world models.

Evaluating world models. These conclusions rest on aggregate measures, such as agreement between policy rankings computed in the model and in the real world (Zhu et al., 2025; Guo et al., 2026; Gao et al., 2026) or action prediction accuracy (Luo et al., 2026; Dyna Robotics, 2026), which confirm practical utility but compound perception, dynamics, and control into a single number. Model quality itself is otherwise reported with metrics inherited from video generation: frame-level SSIM (Wang et al., 2004), PSNR (Mannos & Sakrison, 1974), LPIPS (Zhang et al., 2018) and DreamSim (Fu et al., 2023), or distributional FVD (Unterthiner et al., 2018), which reward perceptual quality rather than accurate dynamics. None of these localize what improved. Very recently, segmentation-based metrics that directly measure structural agreement between predicted and ground-truth videos have been proposed (Bagchi et al., 2026; Pallotta et al., 2026). We adopt this formulation, and report it separately for agents and objects being interacted with.

Learning interaction dynamics from pixels. A parallel line of work questions whether pixel prediction is the right objective for world modeling at all. LeCun et al. (2022) argues that predicting in observation space forces a model to allocate capacity to detail that is unpredictable and irrelevant to action, motivating joint-embedding architectures (Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025). If this critique is correct, capacity should flow to what dominates the frame and is easiest to predict (e.g. the background) rather than to the small, frequently occluded objects where the interaction actually happens. In this work, we test this hypothesis directly at scale.

Several approaches to improving agent and object fidelity have been proposed. On the agent side, skeleton-based conditioning has been explored (Wang et al., 2025b; Wu & Gao, 2026), motivated by cross-embodiment transferability, though its effect on scaling has not been examined. On the object side, earlier work proposed explicit object-centric representations (Ferraro et al., 2025; Feng et al., 2025), or auxiliary generation targets (Yan et al., 2026), however, those sacrifice the scalability of generic diffusion architectures and do not generalize to ego-centric video. PhysisForcing (Zhang et al., 2026b) adds trajectory and semantic alignment losses over physics-informative regions, but is evaluated on third-person robot manipulation with a fixed camera and largely unoccluded interactions. We find it yields no improvement at scale in our setting, and propose an approach that does not rely on those conditions, raising the level object fidelity converges to.

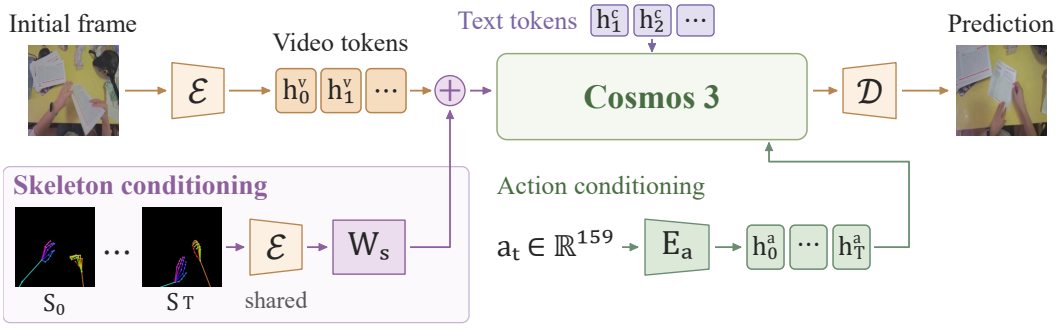

Figure 2: **Cosmos 3 with skeleton conditioning.** The projected skeleton sequence $S_{0:T}$, shown in bottom left, is encoded by the same frozen VAE $\mathcal{E}$ as the video, and its latent patches are projected through a zero-initialized W_s and added to the corresponding video embeddings. The skeleton is derived from the camera intrinsics to better align agent's actions with the visual token grid.

3 Experiment Protocol

We train two variants of the Cosmos 3 video diffusion model (NVIDIA, 2026), differing in both design and size (Section 3.1), each with and without skeleton conditioning, on nested data subsets spanning 300 to 30,000 hours (Section 3.2). All configurations are evaluated on an out-of-distribution benchmark drawn from a separate recording campaign, measuring how accurately the agent's body and the manipulated objects are rendered, with perceptual metrics as a control (Section 3.3). Training details are provided in appendix G.

3.1 World Model Design

We build on the pretrained Cosmos 3 video diffusion model (NVIDIA, 2026) (Figure 2). Given an initial RGB observation x_0, a text description c, and an agent-motion sequence $a_{0:T}$, it predicts future frames $x_{1:T}$. A frozen video autoencoder $\mathcal{E}$ maps the video to spatiotemporal latents, which a dual-tower backbone processes together with text and action tokens: text attends causally, while video and action tokens attend jointly to one another and to the text context. We study the Edge and Nano variants, which differ in size, design, and pretraining. Both share this architecture, but Edge is a 4B model whose 2B denoising transformer is custom-designed and trained from scratch, while Nano is a 16B model whose 8B transformer is adapted from pretrained Qwen3-VL 8B (Bai et al., 2025). They are therefore not two points on a capacity axis, and we treat differences between them as differences between model variants rather than as a scaling trend in model size.

We extend the model's action space with full hand articulation and shoulder/elbow positions:

$$s_t = [\underbrace{c_t}_{9}, \underbrace{w_t^L, w_t^R}_{18}, \underbrace{h_t^L, h_t^R}_{120}, \underbrace{b_t}_{12}] \in \mathbb{R}^{159},$$

where c_t and $w_t^{L/R}$ are camera and wrist poses, each with 3D translation and 6D rotation; $h_t^{L/R}$ each hold 20 hand-joint positions in the corresponding wrist's frame and b_t holds both shoulders and elbows. Actions a_t are obtained by normalizing s_0 and the element-wise differences $s_t - s_{t-1}$; a learned linear encoder maps each to one action token, which remains clean during denoising.

We train the diffusion branch and its conditioning modules with rectified flow matching (Liu et al., 2023), keeping the VLM and VAE fixed. Let $z = \mathcal{E}(x_{0:T})$ be the clean video latents, $\epsilon \sim \mathcal{N}(0, I)$, and M^{pred} a binary mask selecting future latents. At noise level $\sigma \in [0, 1]$,

$$\begin{aligned} z_\sigma &= z + \sigma M^{\text{pred}} \odot (\epsilon - z), \\ \mathcal{L}_{\text{FM}} &= \mathbb{E}\Big[\big\| M^{\text{pred}} \odot (v_\theta(z_\sigma, \sigma; c, a_{0:T}) - (\epsilon - z))\big\|_2^2\Big]. \end{aligned} \tag{1}$$

The initial observation remains clean. At inference, the observed latents are fixed, and we decode the resulting video latents with the VAE decoder $\mathcal{D}$.

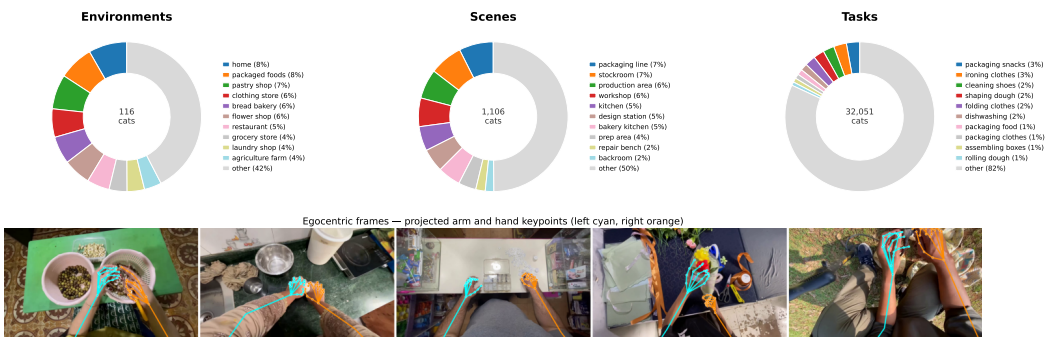

Figure 3: **Dataset overview.** *Top:* distribution of clips over environments, scenes, and tasks (ten most frequent categories each; the gray wedge aggregates the long tail). *Bottom:* representative first-person frames from four environments, overlaid with the dataset's per-frame keypoints.

Skeleton conditioning. Action tokens describe agent motion in a coordinate space of poses and joint angles, leaving the model to learn how it projects onto the pixels. We remove this inference by additionally conditioning on a spatially aligned skeleton sequence (Figure 2, bottom left), placing the agent's geometry on the video-token grid. For each frame we project the two 21-joint hands and their shoulder–elbow–wrist chains onto the image plane: a joint at camera-frame position $q_{t,j}$ has pixel coordinates $u_{t,j} = (K_t q_{t,j})_{1:2}/(K_t q_{t,j})_3$ and depth $d_{t,j} = (q_{t,j})_3$, with camera intrinsics K_t.

We rasterize the joints and connecting bones onto a black RGB canvas. Projection discards depth, so we encode it in colour: hue identifies which finger or limb a primitive belongs to, and value darkens with distance, so the skeleton carries 3D orientation as well as image-plane position. A primitive at depth d_p (in meters, the mean endpoint depth for bones) is drawn in

$$\begin{aligned} \eta(d_p) &= \mathrm{clip}\left(\frac{d_p - 0.15}{0.70 - 0.15}, 0, 1\right), \\ C_p &= \mathrm{HSVtoRGB}\left(\frac{\phi_p}{360}, \alpha_p, 1 - \beta_p \eta(d_p)\right), \end{aligned} \tag{2}$$

where $\eta(d_p)$ maps depth onto a normalized shading factor over the working range, ϕ_p is the hue assigned to the primitive's finger or limb, and (α_p, β_p) control saturation and the strength of depth shading; exact values are given in Appendix G.

To condition the diffusion model, we encode the rasterized sequence $S_{0:T}$ with the same frozen VAE used for video and partition the latents into patches aligned with the video grid. We denote each vectorized skeleton latent patch by z_i^{skel} and its projected token by $h_i^{\mathrm{s}} = W_s z_i^{\mathrm{skel}}$, where W_s is a bias-free linear projection. We add this skeleton token to the corresponding video token as $\widetilde{h}_i = h_i^{\mathrm{v}} + h_i^{\mathrm{s}}$. The skeleton is supplied at both training and inference time.

3.2 Training

We train all models on a large-scale ego-centric dataset of everyday human manipulation: 1.15M first-person clips totalling 30k hours (3.24×10^9 frames at 30 fps, mean clip length 94 s), each recording a manipulation activity from a head-mounted camera. Every frame carries calibrated intrinsics and a 6-DoF camera pose, dense 3D hand keypoints (21 per hand) and wrist poses, upper-body keypoints, and a free-form activity description. Crucially for studying generalization, the corpus is highly diverse: 116 environment categories (homes, bakeries, flower shops, farms, repair shops), 1,106 scene categories and 32,051 tasks from 14,019 participants, with no environment type exceeding 8.3% of the data. Figure 3 summarizes the environment, scene and task distributions as well as visualizes a few samples from the dataset.

We train on a *nested* ladder of subsets at 300, 1k, 3k, 10k and 30k clip-equivalent hours, spanning two orders of magnitude. Each budget is a checkpoint along a progressive pass through a fixed episode order balanced over environment, scene and task, so larger budgets extend smaller ones, composition is preserved at every scale, and differences along the ladder reflect how much data a model has seen rather than which data it received. We report further details in Appendix A.

3.3 EVALUATION

Dataset. Because we ask whether *generic* models of object dynamics can be learned from ego-centric human video, we evaluate out of distribution. Our evaluation set comes from a separate recording campaign: none of its sessions appear in training, and since platform, operators and sites all differ, no location or participant is shared. 37% of its environment categories, accounting for a quarter of its clips, never appear in training (Appendix F). From this evaluation set we hand-curated 150 clips under three criteria: (i) *non-trivial agent–object interaction* — each clip contains active manipulation rather than idle or transit motion; (ii) *segmentable objects* — the manipulated objects are large enough in the image to be segmented accurately, which mask-based metrics rely on; and (iii) *accurate body-joint annotations* — clips with accurate upper-body and hand keypoints only.

Our key metrics are computed over the regions occupied by the hands and the manipulated objects rather than the full frame. We annotate every clip with SAM 2 (Ravi et al., 2025), prompting a few points on the first frame for the hands and each manipulated object and propagating them with the video tracker, then manually verify and correct the resulting tracks. The evaluation set comprises 150 clips of 32 frames at 10 fps with $11{,}328$ instance masks; models are evaluated on 17-frame windows, one observed frame and 16 predicted. We will release the entire evaluation set.

Metrics. We measure structural consistency and perceptual quality. SCS (Bagchi et al., 2026) segments and tracks key scene elements through the predicted and ground-truth videos and averages mask IoU, capturing whether the provided actions are accurately reflected in the generated world; we report it separately for the hands (agent) and the manipulated objects. LPIPS (Zhang et al., 2018) serves as a perceptual control, so that differences in structural fidelity can be checked against differences in image quality.

Because mask tracking is itself imperfect, even a perfect prediction would not score an SCS of 1.0. We estimate achievable maximum by re-initializing the tracker from different frames of the same video, with prompts sampled from the ground-truth masks, and measuring agreement between the resulting tracks: 0.93 for the agent and 0.90 for objects on ground-truth video, with comparable values on generated video, so the bound applies to model outputs too. Both lie well above the scores our models attain, so the differences we report reflect the models rather than the limits of the metric.

4 HOW FAR DOES EGO-CENTRIC DATA TAKE US?

Data scaling saturates agent fidelity. We first ask what data volume alone provides, measuring directly how accurately the model renders the agent's body and the objects it manipulates. Perceptual metrics are reported in the appendix as a control, confirming that the differences we measure are not confounded with image quality. The grey portion of the plots in Figure 4 shows agent and object fidelity as a function of pretraining data for Edge and Nano Cosmos 3 variants (dashed lines). Both improve across the full range, although the gap between modeling the agent and modeling its effect on the world grows with data scaling. Nano achieves both higher absolute fidelity and faster improvement than Edge, most visibly on objects ($+0.14$ versus $+0.10$). Read at face value, this supports the scaling-centric view: fidelity rises with both data and model, suggesting continued scaling might eventually close the gap.

The two axes differ, however, in what the curves reveal about their limits. Agent fidelity is visibly saturating: it gains 0.12 SCS between 300 and 3k hours but only 0.06 between 3k and 30k, and fitting a saturating curve of the form $S_\infty - cD^{-\alpha}$, standard in scaling-law analyses (Kaplan et al., 2020; Hoffmann et al., 2022), places its asymptote at 0.811. Object fidelity shows no comparable deceleration, and a curve without curvature does not constrain its asymptote, so the object limit remains undetermined within our range, which is why we plot these curves only over the observed range. Rather than extending data collection by another order of magnitude, we next use skeleton conditioning as an instrument for locating that limit.

Conditioning fixes the agent, revealing object scaling. Skeleton conditioning supplies the agent's projected body configuration in the camera frame, so the model no longer has to infer it from action tokens. The effect on the agent axis is large (solid curves in Figure 4, left): with 300 hours, the skeleton-conditioned Nano variant reaches the agent fidelity the baseline attains only at roughly 15k hours - a $\sim 50\times$ reduction in data - and gains just 0.06 SCS over the remaining two

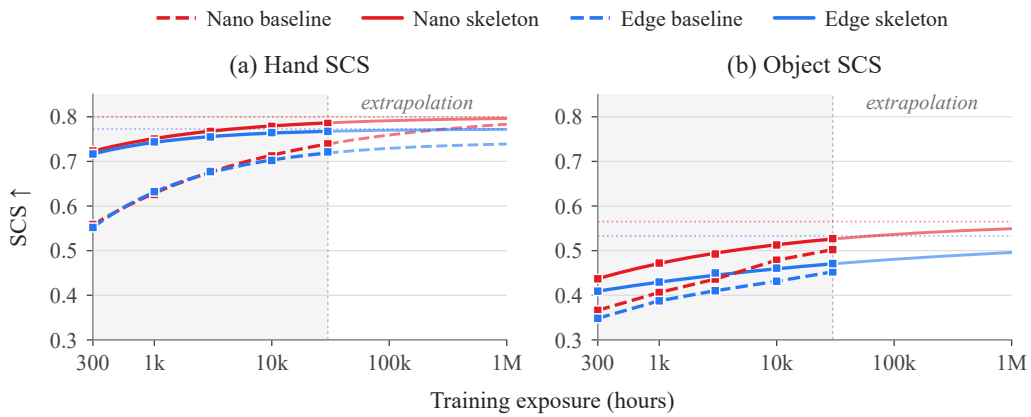

Figure 4: **Scaling reaches the agent's limit, not the object's.** SCS for the agent (left) and manipulated objects (right) versus training data, with and without skeleton conditioning. Skeleton conditioning reaches the agent's limit with far less data, while object fidelity converges far below.

orders of magnitude. Crucially, conditioning changes *when* the agent's limit is reached, not *what* it is: the two designs converge to about 0.01 SCS of one another (0.811 and 0.800). Additional ego-centric video therefore cannot meaningfully improve this axis, and saturating it early gives us an instrument for locating the object limit within our budget.

Objects benefit from conditioning as well (Figure 4, right), the Nano variant reaching at 300 hours the baseline's fidelity at 3k, though this advantage narrows across the ladder. This could indicate a shortcut: by supplying body configuration directly, conditioning might spare the model visual inference that could support interaction modeling. We test this in Appendix C.1 by randomly dropping the skeleton during training, which retains that inference, and find no evidence of a shortcut. Conditioning therefore isolates the quantity we care about: with the agent held near its ceiling across the ladder, the object curve is largely unconfounded by improvements in agent fidelity.

Measured this way, object fidelity improves by only 0.09 SCS across two orders of magnitude, and unlike the baseline it decelerates, gaining 0.056 between 300 and 3k hours but 0.035 between 3k and 30k. The curve is now in its saturating regime and its asymptote is identifiable: 0.565, of which the model has already realized 93% at 30k hours. Normalized by the maximum the metric can register for each category, the agent converges to 86% of its measurable range and objects to 63%. The limitation is therefore not that object dynamics improve slowly with data, but that the level they improve toward lies far below the agent's.

5 Can the object ceiling be moved?

The previous section established a low limit on object fidelity that additional data is unlikely to overcome. We now ask whether changing the supervision can.

Existing interaction-focused supervision does not transfer. PhysisForcing (Zhang et al., 2026b) localizes 'physics-informative' regions using 2D point tracking and monocular depth, then within them trains intermediate diffusion features to support point tracking against CoTracker3 (Karaev et al., 2025) references and aligns their pairwise token-similarity matrix with a frozen V-JEPA 2 encoder (Assran et al., 2025). Re-implementing both objectives, we find they do not improve over the skeleton-conditioned baseline at 10k hours (Table 1). Both the localization heuristic and the tracking supervision rely on a static camera and largely unoccluded objects, and in ego-centric manipulation neither holds: the camera moves with the agent, and the object is occluded by the very hand acting on it at the moment of contact. Next, we design an approach that depends on neither assumption.

Localizing dynamic regions. As shown in Figure 12 (a), we first initialize a regular grid of queries in the first frame and track them through the video with D4RT (Zhang et al., 2026a). Combining each point's 3D displacement and tracking reliability with a soft spatial prior around the projected hands, and expanding the resulting scores over neighbouring latent cells, yields a dynamic region map $M^{\mathrm{dyn}} \in [0,1]^{H \times W}$ (see Appendix D for details).

Variant	Agent ↑	Object ↑	LPIPS ↓
Skeleton	0.779 ± 0.002	0.513 ± 0.004	0.235 ± 0.002
Prior auxiliary objectives			
PhysisForcing	0.780 ± 0.002	0.515 ± 0.005	0.234 ± 0.002
Supervision allocation			
Reweighting	0.781 ± 0.002	0.516 ± 0.004	0.235 ± 0.001
Dynamic noise	0.786 ± 0.001	0.526 ± 0.004	0.234 ± 0.002
Dynamic noise (random)	0.781 ± 0.002	0.517 ± 0.005	0.233 ± 0.002
Ours	**0.788** ± 0.002	**0.532** ± 0.003	0.235 ± 0.002

Table 1: **Regional supervision ablation at 10k hours.** Combining dynamic noise scheduling and loss reweighting achieves the highest object fidelity, without sacrificing perceptual quality. Mean ± SD across five seeds.

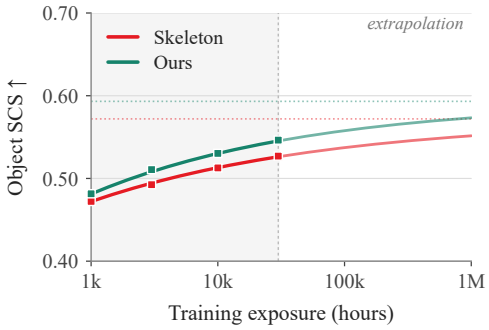

Figure 5: **A higher object ceiling.** Object fidelity across the data ladder, with and without dynamic-region supervision. Ours exceeds the skeleton baseline at every budget, and converges to a higher limit.

Shifting the supervision. Within these regions we raise the noise level, so the denoising target can no longer be satisfied by propagating local appearance and must be predicted from context and dynamics. With $\Omega = \{i : M_i^{\text{pred}} = 1\}$ the predicted latent locations, z_σ from Equation 1 and $v = \epsilon - z$, for $i \in \Omega$ and $0 < \sigma < 1$:

$$\delta_i = \lambda M_i^{\text{dyn}} \min(\sigma, 1-\sigma), \qquad \widetilde{z}_{\sigma,i} = z_{\sigma,i} + \delta_i v_i, \qquad \widetilde{v}_i = \left(1 + \tfrac{\delta_i}{\sigma}\right) v_i, \tag{3}$$

where $\lambda \in [0, 1]$ scales the additional noise in the selected regions. This raises the effective noise level from σ to $\sigma+\delta_i$ while retaining σ as the model's noise-level input, and the paired target satisfies $\widetilde{z}_{\sigma,i} - \sigma\widetilde{v}_i = z_i$, so supervision still points toward the same clean latent. For comparison we also reweight the loss within the same regions (described in Appendix D).

Results. We first report an ablation of different components of our objective at 10k hours in Table 1. Reweighting alone does not meaningfully improve fidelity, whereas raising the noise level does, and is the only intervention we test that improves object fidelity. Replacing M^{dyn} with randomly placed regions of the same extent nearly all of the gain, confirming that the benefit comes from where supervision is shifted rather than from the perturbation itself. Reweighting adds a small further improvement in combination at no extra cost, so we include it in our final configuration. LPIPS is unchanged throughout, confirming that our approach does not sacrifice perceptual quality.

Figure 5 reports object interaction fidelity across the data ladder. Our method exceeds the skeleton baseline at every budget, reaching 0.546 against 0.527 at 30k hours (improvement of 3.7%), and fitting the same saturating form places its asymptote above the baseline's. The gain is, however, roughly constant across budgets: the ceiling is raised, not the rate at which it is approached, and it remains far below the agent's. Improving object fidelity to a usable level will require more substantial changes to how world models are trained.

6 Transfer to Humanoids

Setup. We fine-tune Cosmos 3 Nano on ground-truth demonstrations from 11 tasks from the simulated bimanual benchmark of Yang et al. (2025), conditioning on a single start frame and the 138-dimensional humanoid state trajectory and predicting the following 16 frames (17 total at 10 Hz). We compare three variants sharing the same recipe and evaluation clips: Cosmos 3, from the released weights, which have never seen our ego-centric data; +Human, from the matched checkpoint pre-trained on 30,000 h without skeleton conditioning or our supervision scheme; and Ours, from the checkpoint pre-trained with both, which remain active during fine-tuning. The first pair isolates ego-centric human pre-training, the second our design. All evaluations are performed on EgoVLA policy rollouts rather than demonstrations. Further details provided in Appendix E.

Metrics. We report SCS for hands and manipulated objects separately, using manually verified masks. To assess applicability to downstream robotics evaluation, we additionally measure how well each model's rendered rollouts support the same success judgment as the simulator's: within

Table 2: **Transfer to humanoid manipulation.** SCS on the humanoid benchmark, and agreement with the simulator's rendering on policy-rollout success. Contribution columns give the difference between adjacent variants. Higher is better.

	Variant			Contribution	
Metric	COSMOS 3	+HUMAN	OURS	Human pre-train	Ours
Hand SCS	0.63	0.70	**0.88**	+0.07	+0.17
Object SCS	0.70	0.73	**0.79**	+0.03	+0.06
Success agreement	0.67	0.71	**0.81**	+0.04	+0.10

each rollout we locate the 17-frame window in which the outcome is decided, replay its actions through each variant, and apply the benchmark's success criterion to both renderings, reporting how often the two verdicts agree. Restricting to the critical window isolates world-model fidelity from drift over long rollouts. Appendix E gives the full protocol, and per-task results.

Results Ego-centric human pre-training transfers, improving every measure and most clearly hand SCS ($+0.07$, Table 2). Our conditioning and supervision design contributes more than twice as much on top of it. The pattern matches our human-video findings in one respect and departs from it in another. As there, gains are larger for the agent than for the objects it manipulates ($+0.17$ versus $+0.06$). Unlike there, object fidelity is high in absolute terms for every variant (0.70–0.79), which we attribute to the simulated environment: its objects are rigid and relatively large, so the conditions that make object dynamics hard in ego-centric human video are largely absent. Conclusions about interaction fidelity thus depend heavily on the difficulty of the evaluation setting.

Success agreement follows the same ordering with smaller margins, and is less diagnostic: a single number records that agreement differed without indicating where. SCS localizes the residual error to the objects. Over the prediction horizon, hand SCS for OURS is nearly flat, falling from 0.89 at frame 2 to 0.85 at frame 16, while object SCS falls more than twice as far, from 0.82 to 0.71. The model knows where the hands will be, but fails to correctly predict the consequences of agent's actions in the world. We illustrate this via qualitative analysis in Appendix B.2.

7 DISCUSSION AND LIMITATIONS

Our results are not an argument against collecting ego-centric human video: more data improved both axes at every scale we measured, and pre-training transferred to humanoids. They are an argument against expecting collection alone to close the interaction gap. Agent fidelity approaches its limit well before 30k hours, while the fidelity of the agent's effects on the world converges far below it, even with our supervision scheme.

Two factors plausibly contribute (see Appendix B.1 for qualitative analysis). The first is irreducible uncertainty: how an object responds to contact depends on mass, friction and stiffness, none of which is perfectly recoverable from a single frame. The second is that the quantities that matter most are not observable in pixels at all. An object's geometry is only partially visible, and the parts most relevant to an interaction (those under the hand, at the point of contact) are exactly the parts the hand occludes. Contacts and forces are never visible. These are precisely what a physics simulator represents exactly and can supply as supervision, while remaining poor at the complex dynamics that motivated learned world models in the first place. Each paradigm is strongest where the other is weakest, and recent work distilling simulators into differentiable generative 3D particle models (Wang et al., 2025a; Duisterhof et al., 2026) suggests a path to combining them.

Limitations. Our two Cosmos 3 variants differ not only in size but in pre-training and design, so their differing limits cannot be attributed to capacity alone. We will report results for the 64B variant to address this. Our data ladder spans two orders of magnitude and couples data volume with optimisation budget. It therefore establishes where fidelity is heading under joint scaling, not the effect of either factor alone, and the asymptotes we report hold only insofar as the fitted regime continues beyond 30k hours. Finally, both our training and evaluation sets come from a single collection protocol, and the evaluation set is limited in scale, which constrains how far our conclusions generalize to ego-centric video as a whole.

AI USE STATEMENT

In this work, we used generative AI tools for interpreting results. We have not used generative AI tools for other tasks that require disclosure. Additionally, we used generative AI tools for several tasks with recommended disclosure: editing parts of the paper, creating and modifying figures, editing software code, searching for related literature. We have reviewed all AI-assisted work. Every reported number is computed by scripted evaluation from saved model predictions and all characterizations of prior work were checked against the primary sources. We take responsibility for the final content of this work, including text or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 3.1 describes the backbone, the two model variants as well as the skeleton conditioning, including the projection and colour encoding, and Section 5 the dynamic-region localization and noise-scheduling objective, with the full algorithm and hyperparameters in Appendix D. Section 3.2 describes the training corpus and the nested data ladder, with the sampling convention in Appendix A.4. Section 3.3 describes the evaluation benchmark, its curation criteria and mask annotation procedure, as well as the metrics, including our estimate of the maximum score each category can attain. Section G lists optimization settings, and further training and implementation details. Appendix E gives the humanoid fine-tuning setup and per-task results, and Appendix C.1 the skeleton-dropout control. We will publicly release our code, the pre-trained model checkpoints, and the evaluation benchmark with its instance mask annotations, so that every number we report can be reproduced.

REFERENCES

Brian Acosta, William Yang, and Michael Posa. Validating robotics simulators on real-world impacts. *IEEE Robotics and Automation Letters*, 7(3):6471–6478, 2022.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023.

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

Anurag Bagchi, Zhipeng Bao, Homanga Bharadhwaj, Yu-Xiong Wang, Pavel Tokmakov, and Martial Hebert. Walk through paintings: Egocentric world models from internet priors. In *ECCV*, 2026.

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

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *CVPR*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *TMLR*, 2024.

David Blanco-Mulero, Oriol Barbany, Gokhan Alcan, Adrià Colomé, Carme Torras, and Ville Kyrki. Benchmarking the sim-to-real gap in cloth manipulation. *IEEE Robotics and Automation Letters*, 9(3):2981–2988, 2024.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70M: Captioning 70M videos with multiple cross-modality teachers. In *CVPR*, 2024.

Kenneth J. W. Craik. *The Nature of Explanation*. Cambridge University Press, 1943.

Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. RoboNet: Large-scale multi-robot learning. In *CoRL*, 2019.

Bardienus P Duisterhof, Kaifeng Zhang, Adam Hung, Bowen Wen, Stan Birchfield, Yunzhu Li, Deva Ramanan, and Jeffrey Ichnowski. PointZero: 3D point track completion for learning transferable 3d dynamics. *arXiv preprint arXiv:2609.19142*, 2026.

Dyna Robotics. Dyna-2: A 1-million-hour scaling law for world-action models. Technical report, 2026. `https://www.dyna.co/dyna-2`.

Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*, 2018.

Fan Feng, Phillip Lippe, and Sara Magliacane. Learning interactive world model for object-centric reinforcement learning. In *NeurIPS*, 2025.

Stefano Ferraro, Pietro Mazzaglia, Tim Verbelen, and Bart Dhoedt. FOCUS: Object-centric world models for robotic manipulation. *Frontiers in Neurorobotics*, 2025.

Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *ICRA*, 2017.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *NeurIPS*, 2023.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, et al. DreamDojo: A generalist robot world model from large-scale human videos. In *ICML*, 2026.

Yanjiang Guo, Lucy Shi, Jianyu Chen, and Chelsea Finn. Ctrl-World: A controllable generative world model for robot manipulation. In *ICLR*, 2026.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *NeurIPS*, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *ICML*, 2019.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. In *ICCV*, 2025.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. DROID: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Hao Luo, Wanpeng Zhang, Yicheng Feng, Sipeng Zheng, Haiweng Xu, Chaoyi Xu, Ziheng Xi, Yuhui Fu, and Zongqing Lu. Being-H0.7: A latent world-action model from egocentric videos. *arXiv preprint arXiv:2605.00078*, 2026.

James Mannos and David Sakrison. The effects of a visual fidelity criterion of the encoding of images. *IEEE Transactions on Information Theory*, 20(4):525–536, 1974.

Michaël Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In *ICLR*, 2016.

NVIDIA. Cosmos 3: Omnimodal world models for physical AI. *arXiv preprint arXiv:2606.02800*, 2026.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open X-Embodiment: Robotic learning datasets and RT-X models: Open X-Embodiment collaboration. In *ICRA*, 2024.

Enrico Pallotta, Sina Mokhtarzadeh Azar, Lars Doorenbos, Serdar Ozsoy, Umar Iqbal, and Juergen Gall. EgoControl: Controllable egocentric video generation via 3D full-body poses. In *CVPR*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

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

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. In *ICLR*, volume 2025, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Jürgen Schmidhuber. *Making the world differentiable: on using self supervised fully recurrent neural networks for dynamic reinforcement learning and planning in non-stationary environments*, volume 126. Inst. für Informatik, 1990.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using LSTMs. In *ICML*, 2015.

Richard S Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin*, 2(4):160–163, 1991.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Ruben Villegas, Arkanath Pathak, Harini Kannan, Dumitru Erhan, Quoc V Le, and Honglak Lee. High fidelity video prediction with large stochastic recurrent neural networks. *NeurIPS*, 2019.

Chen Wang, Chuhao Chen, Yiming Huang, Zhiyang Dou, Yuan Liu, Jiatao Gu, and Lingjie Liu. PhysCtrl: Generative physics for controllable and physics-grounded video generation. *NeurIPS*, 2025a.

Yuang Wang, Chao Wen, Haoyu Guo, Sida Peng, Minghan Qin, Hujun Bao, Xiaowei Zhou, and Ruizhen Hu. Precise action-to-video generation through visual action prompts. In *ICCV*, 2025b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Zhuoyuan Wu and Jun Gao. OSCAR: Omni-embodiment action-conditioned world model for robotics. *arXiv preprint arXiv:2606.04463*, 2026.

Haodong Yan, Hang Yu, Zhide Zhong, Weilin Yuan, Xin Gong, Zehang Luo, Chengxi Heyu, Junfeng Li, Wenxuan Song, Shunbo Zhou, and Haoang Li. Open-world hand-object interaction video generation based on structure and contact-aware representation. In *CVPR*, 2026.

Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. EgoVLA: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025.

Chuhan Zhang, Guillaume Le Moing, Skanda Koppula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle K Barral, Raia Hadsell, et al. Efficiently reconstructing dynamic scenes one D4RT at a time. In *CVPR*, 2026a.

Peiwen Zhang, Yufan Deng, Shangkun Sun, Juncheng Ma, Duomin Wang, Jonas Du, Zilin Pan, Ye Huang, Hao Liang, Songyan Huang, Ruihua Zhang, Enze Xie, Ming-Yu Liu, and Daquan Zhou. PhysisForcing: Physics reinforced world simulator for robotic manipulation. *arXiv preprint arXiv:2606.28128*, 2026b.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, et al. EgoScale: Scaling dexterous manipulation with diverse egocentric human data. *arXiv preprint arXiv:2602.16710*, 2026.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. IRASim: A fine-grained world model for robot manipulation. In *ICCV*, 2025.

A Training Set Details

A.1 Overview and modalities

Our training set is a large-scale collection of egocentric video of humans performing everyday manipulation activities, recorded from a head-mounted camera. Each clip is a single continuous recording (mean length $94.3\,\mathrm{s}$, median $94.5\,\mathrm{s}$). We use the *full* corpus of $1{,}146{,}100$ clips ($30{,}012$ hours, 3.24×10^{9} RGB frames at $30\,\mathrm{fps}$). Every clip is annotated per frame with the following aligned modalities:

- **RGB** egocentric video (undistorted pinhole projection).
- **Camera calibration**: per-frame intrinsics and a 6-DoF camera pose (camera-to-world, metric scale) recovered by SLAM.
- **Hands**: 3D keypoints for both hands (21 landmarks each) and 6-DoF wrist poses, expressed in the camera frame.
- **Body**: upper-body keypoints in a SMPL-H–compatible layout.
- **Language**: a task label and a free-form natural-language description of the activity.

A.2 Scale and diversity

Table 3 reports headline statistics. The corpus is diverse along every axis we measured: 116 environments, $1{,}106$ scenes, $32{,}051$ tasks (after synonym normalization; see below), and $14{,}019$ participants. Importantly, coverage is not concentrated in a few dominant categories. The most common environment (*home*) accounts for only 8.3% of clips; reaching 80% coverage requires 22 of the 116 environments, 53 scenes, and roughly $1{,}170$ tasks. Participation is likewise long-tailed: the median participant contributes 19 clips while the most prolific contributes $5{,}019$, and the top 1% of participants account for 17% of all clips (Fig. 6, right).

Table 3: Training set summary statistics (full corpus).

Statistic	Value
Clips (recordings)	1,146,100
Total duration	30,012 hours
Total frames (30 fps)	3,241,303,112
Mean / median clip length	94.3 s / 94.5 s
Clip length range	2.6–240.0 s
Unique environments	116
Unique scenes (merged)	1,106
Unique tasks (merged)	32,051
Unique participants	14,019
Clips / participant (med., max)	19, 5,019

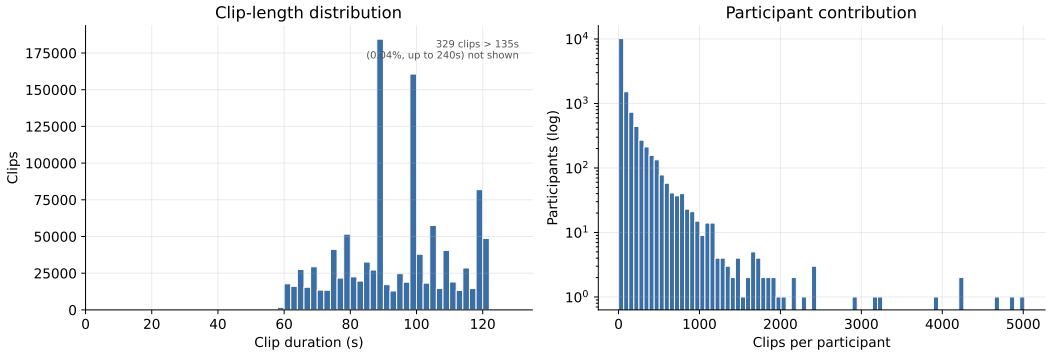

Figure 6: **Left:** clip-length distribution (mean 94.3 s, median 94.5 s; 99.9% of clips are $\leq$ 123 s). The axis is truncated at 135 s; a sparse tail (0.04% of clips) extends to a 240 s cap and is omitted. **Right:** histogram of clips per participant (log y-axis), showing a heavy tail across 14,019 contributors.

A.3 Category distributions

Figure 7 shows the 30 most frequent environments and scenes, and Figure 8 decomposes tasks into actions and manipulated objects. The environment mix spans domestic settings (homes, kitchens, bedrooms), food production (bakeries, pastry and bread shops, restaurants), retail (clothing, grocery, gift, and hardware stores), and craft/repair workshops (woodworking, sewing, shoe and motorcycle repair), among others. Tasks are dominated by fine bimanual manipulation such as ironing and folding clothes, shaping and rolling dough, cleaning, and packaging, consistent with the free-form descriptions in which *small*, *plastic*, *placing*, *folding*, and *cutting* are among the most frequent content words.

A.4 Data scaling ladder

The scaling experiments use a frozen progressive training manifest and a fixed balanced episode order. Each consumed manifest clip contributes its first 17-frame window, starting at frame 0, exactly once. Larger budgets therefore extend the same training prefix. These manifest clips are the short training units extracted from the source recordings, rather than the full recording sessions counted in the dataset overview.

We report data exposure in *clip-equivalent hours*, using the normalization recorded in the frozen training plan. If $N = 33{,}512{,}693$ is the number of manifest clips and n clips have been consumed, the recorded exposure is

$$H(n) = 30{,}000\,\frac{n}{N}\ \text{hours}. \tag{4}$$

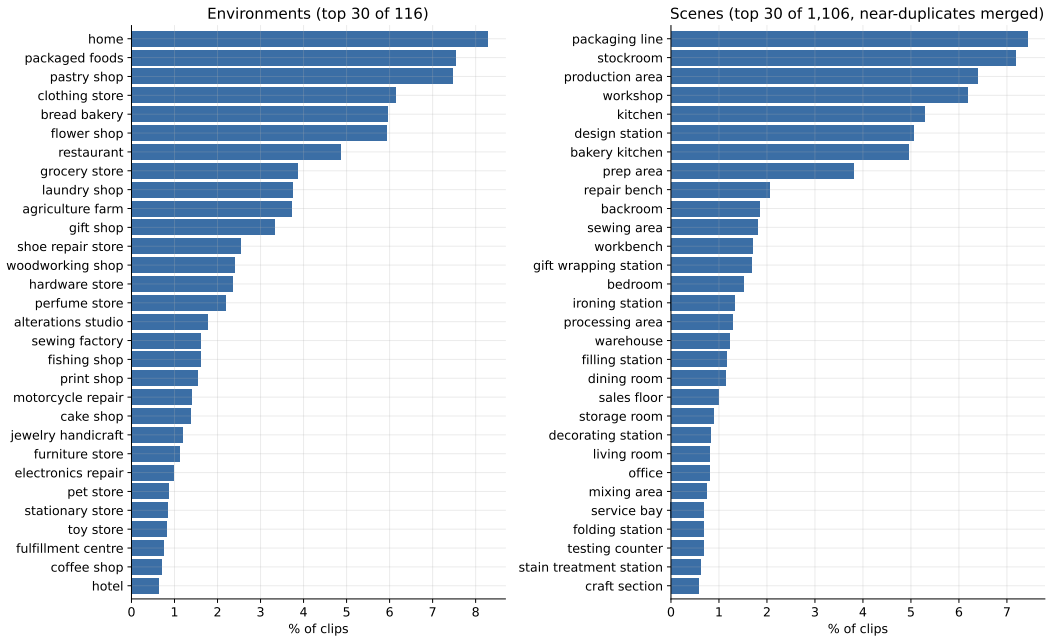

Figure 7: Most frequent **environments** (left, top 30 of 116) and **scenes** (right, top 30 of 1,106, near-duplicates merged), as a percentage of all clips.

Table 4: **Progressive one-pass data budgets.** Each row is a checkpoint on the same trajectory for a given model configuration. Exposure and optimization compute increase together.

Nominal hours	Recorded hours	Optimizer steps	Consumed clips
300	300.209	655	335,360
1,000	1,000.084	2,182	1,117,184
3,000	2,999.795	6,545	3,351,040
10,000	9,999.927	21,818	11,170,816
30,000	29,999.781	65,454	33,512,448

This assigns the full manifest the nominal 30,000-hour budget; it is a data-budget convention, not the sum of the durations of the sampled 17-frame windows. With global batch size 512, checkpoint s consumes $n = 512s$ manifest clips. Table 4 gives the resulting budgets; the small differences from nominal hours arise from integer optimizer steps. The final checkpoint leaves 245 manifest clips unused.

All ladder comparisons use these prescribed checkpoints and the evaluation protocol in Section G. Missing evaluations remain unreported; neither nearby training checkpoints nor interpolated scores are substituted for a ladder measurement.

B Qualitative analysis

B.1 Human samples

Figure 9 shows an example of improvement of our model over vanilla Cosmos 3 (top) and four remaining failure modes (bottom). Though our method improves the agent fidelity and some aspects of object interaction, as illustrated by the package handling sample, we still observe several failure modes. Specifically, contact modeling remains challenging, as illustrated by spoon example, where only one of the two objects is moved by the agent's right hand. The model also struggles with identify physical properties of the objects from RGB frames, as shown on the example of the rigid board being interpreted as stretchable. In addition, object permanence and non-rigid object dynamics

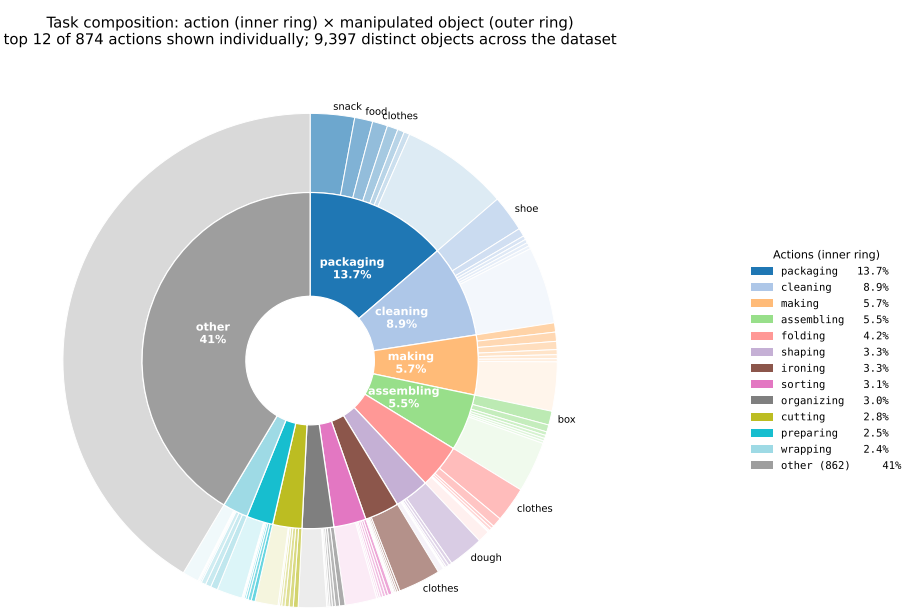

Figure 8: **Task composition.** Each task label decomposes into an action (inner ring) applied to a manipulated object (outer ring); together the corpus spans 874 distinct actions and 9,397 distinct objects (32,051 action–object tasks). For legibility the inner ring shows the 12 most frequent actions with wedge angles proportional to their share of all clips; the remaining 862 actions (41% of clips) are grouped into the grey "other" wedge. The outer ring shows the top objects within each action, with that action's remaining objects aggregated into a lighter "other" sub-wedge. While a few actions (packaging, cleaning, making) are common, each fans out over a broad range of objects.

modeling remain a challenge, as can be seen from the disappearing cucumber sample and incorrect paper folding dynamics. More examples are provided on the supplementary web page, where they are also best viewed in a video form.

B.2 Humanoid samples

Figure 10 compares simulator rollouts with skeleton-conditioned predictions. A successful unloading example is followed by three cases in which the object's response differs from the simulator: a box does not turn under a push, a laptop does not open as the hand rises, and a can does not drop as the fingers open. Thus, accurate agent modeling does not always result in a faithful prediction of the manipulated object even in the simplified simulator environment. More examples are provided on the supplementary web page, where they are also best viewed in a video form.

C Additional Experiments

C.1 Skeleton Dropout Ablation

We test whether always supplying the projected skeleton limits object learning by reducing the model's reliance on visual inference. We compare Cosmos 3 Nano with standard skeleton conditioning against a variant trained with skeleton dropout. For each training clip, we independently set the entire encoded skeleton sequence to zero with probability 0.25, after the frozen VAE and before the learned projection. A single dropout decision applies to all frames and spatial locations in the clip. This intervention leaves the action tokens, observed RGB frame, and video targets unchanged. Both variants receive the full skeleton sequence during evaluation.

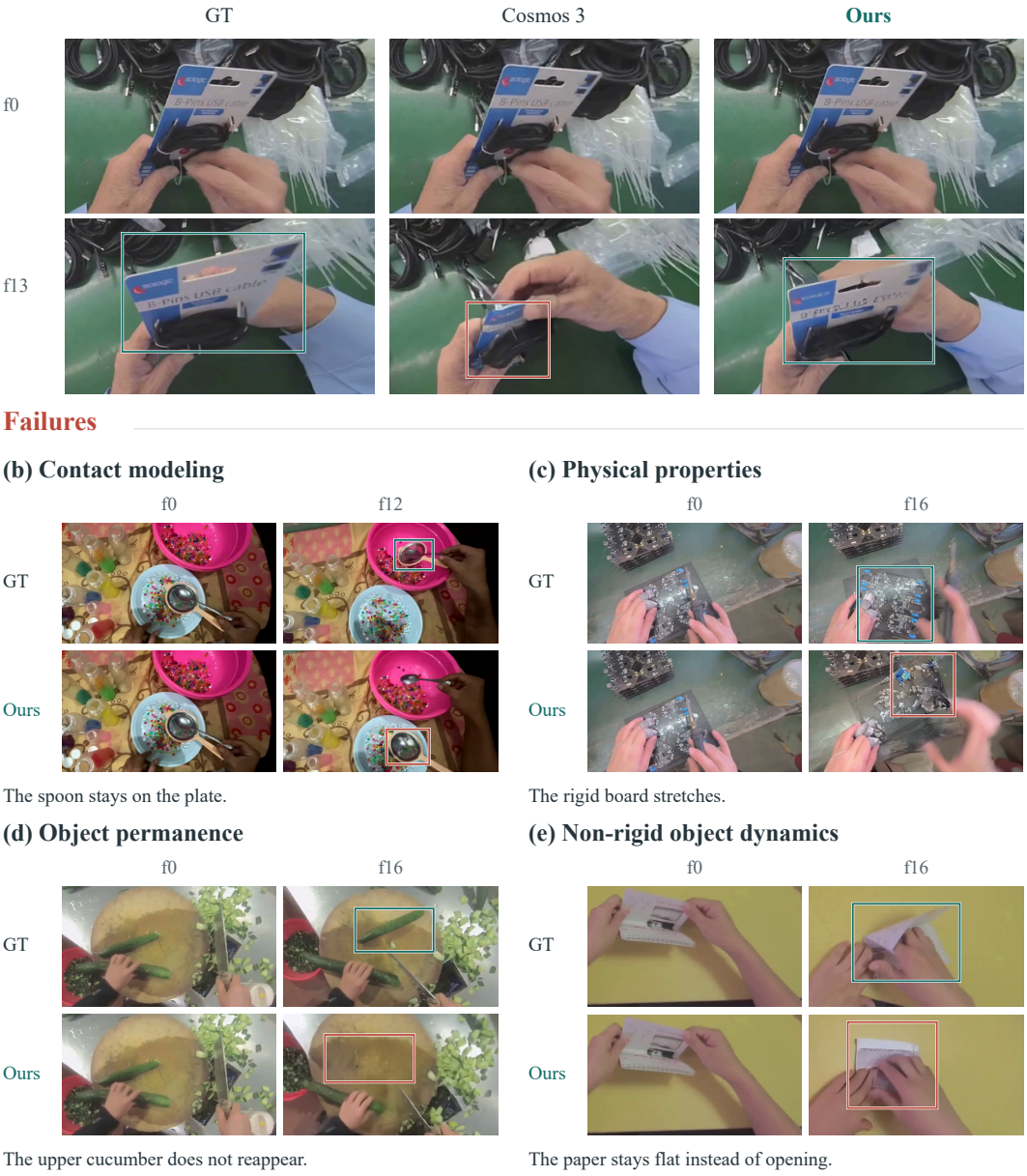

Figure 9: **Improvement and remaining failures in human manipulation.** (a) Our model correctly captures the right hand's motion in the video, whereas vanilla Cosmos 3 fails. (b) The spoon is not lifted. (c) The rigid board stretches. (d) The upper cucumber is missing after the hand moves away. (e) The paper does not open. Frames and crops are matched across methods; f0 is the observed frame. Outlines mark the compared regions. Best viewed on the supplementary web page.

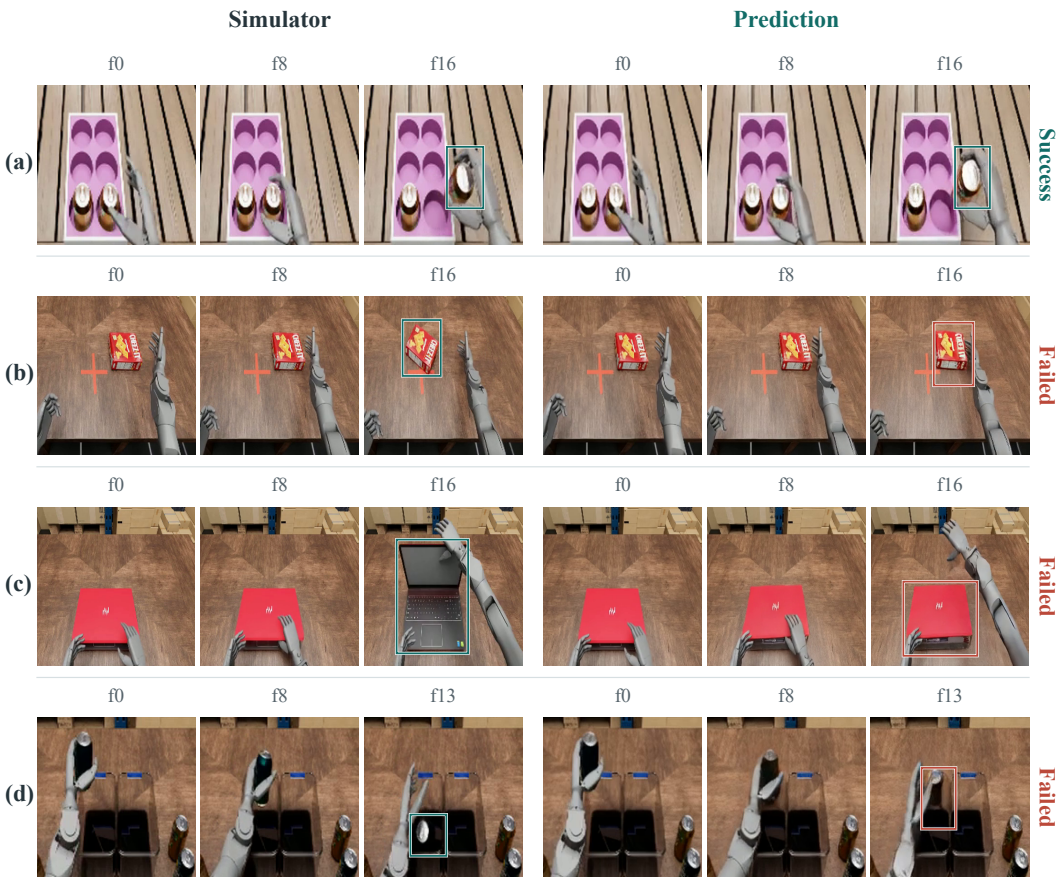

Figure 10: **One successful humanoid rollout and three response failures.** Each row shows one example, with simulator frames on the left and the corresponding predictions on the right. (a) Both rollouts unload the can within the shown window. (b) The simulator's box turns, while the predicted box stays in place. (c) The simulator opens the laptop, while the prediction leaves it closed. (d) The simulator's can drops toward the bin, while the predicted can remains high near its edge. Frame indices are aligned across each pair, with a common crop throughout. Right-hand labels indicate whether the prediction reproduces the simulator's object response within the shown window. Best viewed on the supplementary web page.

We compare checkpoints at 10k training hours, with the same pretrained initialization, training-data schedule, global batch size of 512, and optimizer settings. Both runs use standard denoising supervision without the dynamic-region modifications.

Table 5: **Skeleton dropout ablation.** Cosmos 3 Nano at 10,000 training hours. Dropout is applied only during training; both variants use full skeleton conditioning at evaluation. Higher SCS is better.

Training condition	Agent SCS ↑	Object SCS ↑
Skeleton conditioning	0.779	0.513
+ 25% skeleton dropout	0.778	0.509

As shown in Table 5, the random dropout shows no meaningful effect on agent fidelity and marginally lowers object fidelity. At the matched 10k-hour budget, we therefore observe no scaling benefits from it, lifting the potential concerns of a learning shortcut introduced by skeleton conditioning.

C.2 Perceptual scaling

Figure 11 reports LPIPS (AlexNet) for the configurations in Figure 4 on the final validation set, averaged over the 16 predicted frames and excluding the observed frame.

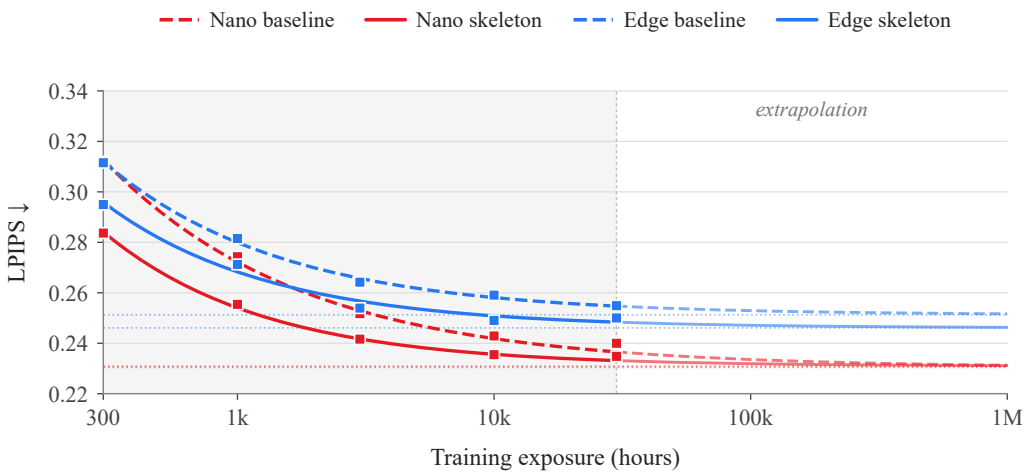

Figure 11: **Perceptual quality across the data-scaling ladder.** LPIPS (lower is better) versus training exposure.

D Dynamic-Region Supervision Details

This appendix details the dynamic-region map M^{dyn} and the two operations it controls in Section 5. We compute the former by first measuring displacement in a common 3D reference frame, where camera motion is factored out and only genuine scene motion remains, and weighting that signal with a soft spatial prior around the projected hands, which the skeleton conditioning already supplies at no additional cost. The result is a per-location score, high where something is both moving in the world and near the agent's hands, which we then use either to raise the local noise level or to reweight the local loss.

Point tracks and motion scores. We initialize 256 queries on a regular 16×16 grid in the first frame. The video VAE encodes the initial observation as one latent frame and temporally compresses the subsequent frames by a factor of four. Thus, a 17-frame video produces $1 + (17 - 1)/4 = 5$ latent frames. We align tracking outputs to latent frame k using video frame $f(k) = 4k$, for $k \in$

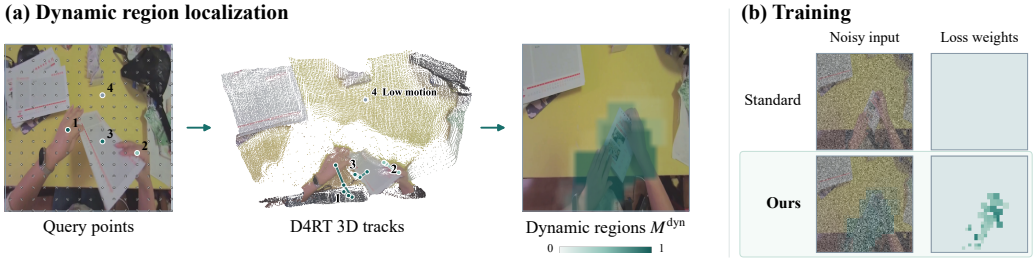

Figure 12: **Dynamic-region supervision. (a)** D4RT tracks queried points in 3D to separate scene motion from camera motion. We combine 3D motion, tracking reliability, and a soft spatial prior around the hands to obtain the dynamic region map M^{dyn}. **(b)** Dynamic-region supervision combines dynamic noise scheduling and loss reweighting in selected regions.

$\{0, \ldots, 4\}$, corresponding to video frames $0, 4, 8, 12, 16$. For query n, D4RT provides an image position $\mathbf{p}_{kn}$ and a 3D displacement ΔX_{kn} from the source frame, measured in a common 3D reference. We multiply tracking confidence and visibility to obtain reliability $\gamma_{kn} \in [0, 1]$. Queries are invalid if any position, confidence, visibility, or displacement is non-finite, the image position is out of bounds, or confidence is below 0.1. For valid queries,

$$m_{kn} = \gamma_{kn} \min\left(\frac{\|\Delta X_{kn}\|_2}{\tau}, 1\right), \qquad \tau = 0.01. \tag{5}$$

Invalid queries receive zero score. The saturation scale τ uses the tracker's native 3D coordinate units.

Hand-proximity prior. Let $\mathcal{H}_{f(k)}$ contain the valid projected hand joints and bone segments in frame $f(k)$. We define

$$\begin{aligned} d^{\text{hand}}_{kn} &= \text{dist}(\mathbf{p}_{kn}, \mathcal{H}_{f(k)}), \\ \rho_{kn} &= \text{sigmoid}\left(\frac{d^{\text{hand}}_{kn} - 8}{4}\right) \text{sigmoid}\left(\frac{48 - d^{\text{hand}}_{kn}}{4}\right). \end{aligned} \tag{6}$$

Distances are measured in pixels on the 512×512 tracking canvas. This prior favours a soft band around the hand geometry, tapering near the skeleton primitives and beyond the hand neighbourhood. It is a spatial prior, not a contact label. We set $\rho_{kn} = 0$ for non-finite positions or absent valid hand geometry; a bone segment is valid when both endpoints are valid and its length is nonzero.

Dynamic-region map. For an image of size $H_I \times W_I$ and latent grid of size $H_z \times W_z$, the continuous latent coordinates of a tracked point are

$$\bar{\mathbf{p}}_{kn} = \left(\frac{W_z}{W_I}(p^x_{kn} + \tfrac{1}{2}) - \tfrac{1}{2},\ \frac{H_z}{H_I}(p^y_{kn} + \tfrac{1}{2}) - \tfrac{1}{2}\right). \tag{7}$$

Its bilinear contribution to latent cell $\mathbf{r} = (r_x, r_y)$ is

$$\omega_{kn}(\mathbf{r}) = \kappa(r_x - \bar{p}^x_{kn})\kappa(r_y - \bar{p}^y_{kn}), \qquad \kappa(a) = \max(0, 1 - |a|). \tag{8}$$

Each query contributes to at most four cells; non-finite positions and out-of-grid contributions are discarded. We multiply the motion score by the hand prior, aggregate overlapping contributions by their maximum, and expand the sparse support:

$$\begin{aligned} B_k(\mathbf{r}) &= \max_n \{\omega_{kn}(\mathbf{r}) m_{kn} \rho_{kn}\}, \\ M^{\text{dyn}}_k(\mathbf{r}) &= M^{\text{pred}}_k(\mathbf{r}) \, \text{clip}_{[0,1]}\left(\frac{\max_{\|\mathbf{r}' - \mathbf{r}\|_\infty \le 2} B_k(\mathbf{r}')}{0.1}\right). \end{aligned} \tag{9}$$

Cells without contributions have $B_k(\mathbf{r}) = 0$. The neighbourhood maximum is 5×5 max pooling with zero padding. The prediction mask M^{pred} sets the map to zero on the conditioning frame. We write $i = (k, \mathbf{r})$ and $\Omega = \{i : M^{\text{pred}}_i = 1\}$ for the predicted locations; all spatial weights are shared across latent channels.

Dynamic-region supervision. In Equation 3, $\lambda \in [0, 1]$ scales the local noise increment δ_i: $\lambda = 0$ gives $\delta_i = 0$, while larger values increase δ_i for fixed σ and $M_i^{\mathrm{dyn}} > 0$. We use $\lambda = 0.65$; for example, at $M_i^{\mathrm{dyn}} = 1$ and $\sigma = 0.5$, the increment is $\delta_i = 0.325$, giving an effective local noise proportion of 0.825 while the model's noise-level input remains 0.5. For $i \in \Omega$, $0 < \sigma < 1$, and $v_i = \epsilon_i - z_i$, the update can be written as

$$\begin{aligned} \delta_i &= \lambda M_i^{\mathrm{dyn}} \min(\sigma, 1 - \sigma), \\ \widetilde{z}_{\sigma,i} &= (1 - \sigma - \delta_i) z_i + (\sigma + \delta_i)\epsilon_i, \\ \widetilde{v}_i &= \left(1 + \frac{\delta_i}{\sigma}\right) v_i. \end{aligned} \tag{10}$$

Because $M_i^{\mathrm{dyn}}, \lambda \in [0, 1]$, the local noise proportion satisfies $\sigma \le \sigma + \delta_i \le 1$. Input and target are modified together: the model still receives the nominal level σ, and the target preserves the clean latent,

$$\widetilde{z}_{\sigma,i} - \sigma \widetilde{v}_i = z_i + (\sigma + \delta_i) v_i - (\sigma + \delta_i) v_i = z_i. \tag{11}$$

The conditioning frame is left unmodified and excluded from the loss.

Reweighting the loss. As an alternative to modifying the noise level, the same map can be used to reweight the loss within the dynamic regions. With gain $g \ge 0$,

$$\begin{aligned} w_i &= \frac{1 + g M_i^{\mathrm{dyn}}}{\langle 1 + g M^{\mathrm{dyn}} \rangle_\Omega}, \\ \mathcal{L}_{\mathrm{reg}} &= \mathbb{E}_{z,\epsilon,\sigma}\left[\sum_{i \in \Omega} w_i \, \| v_\theta(\widetilde{z}_\sigma, \sigma; c, a_{0:T})_i - \widetilde{v}_i \|_2^2 \right], \end{aligned} \tag{12}$$

where $\langle \cdot \rangle_\Omega$ is the mean over predicted locations within a clip, so that $\langle w \rangle_\Omega = 1$ and emphasis is redistributed without changing the average weight. We use $g = 2$. Locations with $M_i^{\mathrm{dyn}} = 0$ retain a positive baseline weight. When the entire map is zero, $w_i = 1$ and both operations reduce to standard training.

Random-region baseline. For the *Dynamic noise (random)* baseline in Table 1, we replace the dynamic-region map with a randomly placed rectangle on the latent spatial grid. For each clip, we sample its center uniformly over the grid and its aspect ratio in normalized coordinates as $\log r \sim \mathcal{U}(\log 0.5, \log 2)$. With a fixed area fraction $a = 0.15$, its width and height are

$$w = W_z \sqrt{ar}, \qquad h = H_z \sqrt{a/r}. \tag{13}$$

If part of the rectangle extends beyond the left or right edge of the latent grid, that part reappears at the opposite edge; the same rule applies to the top and bottom edges. We treat each spatial location in the latent grid as a unit square. Its mask value is the fraction of that square covered by the rectangle: 1 for full coverage, 0 for no coverage, and a fractional value for partial coverage. This preserves a mean mask value of 0.15 on each predicted frame, including for rectangles crossing a boundary. We repeat the same spatial mask across all predicted latent frames and set it to zero on the conditioning frame. Region selection is independent of image content, hand geometry, and tracking outputs. We apply dynamic noise scheduling with $\lambda = 0.65$ and uniform loss weights ($g = 0$).

E Humanoid Evaluation Details

Fine-tuning setup. The humanoid state space is 138-dimensional against the pre-trained checkpoints' 160, so we re-initialize the action projection and leave the rest of the model unchanged. All three variants are fine-tuned on ground-truth demonstrations from the same 11 tasks with the hyperparameters of Section G, with a global batch size of 64. For Ours, skeleton conditioning and dynamic-region supervision remain active during fine-tuning, with the skeleton obtained from the humanoid's forward kinematics.

Evaluation data and masks. SCS is measured on the benchmark's validation split: 72 clips from trained tasks and 48 from held-out ones. Hand masks come from SAM2 (Ravi et al., 2025) prompted at frame 0 with 18 forward-kinematics-projected links per hand and then tracked; object masks come from SAM3 with a per-task text prompt. All masks were manually verified. Success agreement is measured on a separate set of hand-curated windows, so the two measures agree on the ordering of the variants while being computed over different clips.

Table 6: SCS at fixed iterations, averaged over all 11 tasks. The ego-centrically pre-trained variants are converged throughout this range; COSMOS 3 is not, which is why Table 2 selects per variant rather than fixing a shared iteration.

metric	variant	2,000	3,000	9,000
hand SCS	OURS	0.88	0.88	**0.88**
	+HUMAN	0.70	0.70	0.69
	COSMOS 3	0.44	0.56	0.64
object SCS	OURS	**0.77**	0.77	0.76
	+HUMAN	0.70	0.71	0.72
	COSMOS 3	0.56	0.67	0.67

Table 7: Per-task SCS on the humanoid benchmark, at the checkpoints reported in Table 2.

	hand SCS			object SCS		
task	COSMOS 3	+HUMAN	OURS	COSMOS 3	+HUMAN	OURS
Close-Drawer	0.60	0.65	0.88	0.92	0.86	0.97
Flip-Mug	0.69	0.75	0.90	0.70	0.77	0.79
Insert-Cans	0.59	0.64	0.86	0.60	0.61	0.62
Open-Drawer	0.53	0.65	0.86	0.58	0.49	0.62
Pour-Balls	0.71	0.77	0.89	0.75	0.78	0.81
Push-Box	0.65	0.79	0.89	0.81	0.86	0.91
Stack-Can	0.72	0.74	0.87	0.55	0.54	0.70
Unload-Cans	0.66	0.64	0.90	0.67	0.55	0.74
Open-Laptop	0.62	0.72	0.92	0.77	0.89	0.92
Sort-Cans	0.68	0.75	0.86	0.67	0.84	0.87
Stack-Can-Into-Drawer	0.47	0.61	0.82	0.64	0.83	0.75
mean	0.63	0.70	0.88	0.70	0.73	0.79

Checkpoint selection. Each variant is reported at its best checkpoint within a capped window rather than over its whole run, because the three converge at very different rates (Table 6): both ego-centrically pre-trained variants are at their plateau by iteration 2,000, while COSMOS 3 gains 0.19 hand SCS between 2,000 and 9,000 on trained tasks against 0.003 for OURS. Reading the comparison at a single early iteration would therefore measure convergence rate rather than the effect of pre-training. We cap OURS and +HUMAN at 3,000 and COSMOS 3 at 9,000. Selecting on the clips we report is optimistic in absolute terms, but it is applied identically to all three variants; at fixed iterations the two pre-trained variants differ from their selected values by at most 0.005 SCS.

Per-task results. Table 7 reports SCS for each of the 11 tasks and all three variants.

Success-agreement protocol. From each EgoVLA rollout we extract the 17 frames in which the task outcome is decided, e.g. the moment an object is released, a target is reached, or the simulator records success. We replay that clip's actions through each variant and render the prediction. A reviewer then records a success or fail verdict from the rendered clip, once for the simulator's rendering and once for each variant's prediction, applying the benchmark's criterion in both cases, and we report how often the two verdicts agree.

The verdict is therefore about the window, not the episode. This matters because a 17-frame window covers a single interaction event, so a failed episode can contain a cleanly executed sub-task and a successful one a failure. Using the simulator's episode-level flag as the reference would mislabel a substantial fraction of windows. We use that flag only to stratify which clips are reviewed, so the reviewed subset is not skewed toward whichever outcome dominates a task.

Seven of the eleven tasks have such a decisive moment: CLOSE-DRAWER, INSERT-CANS, OPEN-LAPTOP, POUR-BALLS, PUSH-BOX, SORT-CANS and STACK-CAN. In the other four the outcome unfolds over more frames than the clip contains like in OPEN-DRAWER and UNLOAD-CANS; or the success conditions defined in the simulator are too lenient like in FLIP-MUG. Therefore we exclude

Table 8: Agreement with the simulator's rendering on the seven eligible tasks. *Agreement* is the fraction of clips on which the two verdicts coincide; recalls are conditioned on the verdict for the simulator's rendering.

variant	success	agreement	fail recall	success recall
simulator rendering	0.86	—	—	—
COSMOS 3	0.67	0.67	0.50	0.69
+HUMAN	0.67	0.71	0.67	0.72
OURS	0.76	**0.81**	0.67	**0.83**



Figure 13: **Rectifying the evaluation set onto the training camera model.** *Left:* the raw capture from the stereo rig's left camera, a wide-angle double-sphere fisheye (1920×1200); the green box marks the region that maps into the rectified image. *Middle:* that region undistorted to a 908×512 pinhole. *Right:* a training frame, shown for comparison.

these tasks from our evaluation. The review covers six clips per task, 42 in total, with one reviewer and one seed per variant (detailed breakdown provided in Table 8).

F EVALUATION DETAILS

Our evaluation sequences are captured with a multi-camera stereo platform, whereas our training dataset is monocular. We use the stereo left camera for evaluation, however, it is delivered as a double-sphere fisheye. To match the training set sensor, for each frame we unproject the target 908×512 pinhole pixels to rays and reproject them through the double-sphere model, resampling the source image (Fig. 13); the intrinsics are held fixed across all clips. This yields a pinhole camera model consistent with the training data, isolating content/appearance domain shift from any lens-model mismatch.

G IMPLEMENTATION DETAILS

We use 17-frame RGB sequences at 10 fps and 512×512 resolution: one observed frame followed by 16 predicted frames, spanning 1.6 s. We use AdamW with global batch size 512, learning rates of 2×10^{-4} for video-generation and skeleton-projection parameters and 10^{-3} for the action interface, $(\beta_1, \beta_2) = (0.9, 0.99)$, $\epsilon = 10^{-6}$, and zero weight decay. Learning rates are constant, with no warmup. Each configuration follows one progressive pass through the training manifest. The five data budgets are checkpoints along this trajectory, so optimization budget scales with data exposure. Appendix A.4 specifies the sampling convention and checkpoint budgets.

Action normalization uses per-coordinate 1st and 99th percentiles, estimated separately for absolute states and consecutive-frame differences. We reuse the hand-and-camera statistics computed from training videos. For the additional shoulder and elbow coordinates, we randomly sample 5,000 training clips and up to four 17-frame windows per clip, discard invalid windows, and collect their initial states and successive differences. We affinely map each percentile interval to $[-1, 1]$ without clipping, using the absolute-state statistics for the initial action token and the delta statistics for subsequent tokens. These statistics remain fixed throughout training and evaluation.

We evaluate EMA checkpoints on the same 150 held-out clips, using the rectified left view of the stereo evaluation recordings. Generation uses UniPC with 35 sampling steps, classifier-free guid-

ance 5, and seed 12345 plus the validation-clip index, fixed across checkpoints. Agent and object SCS exclude the observed frame and average over future frames 1–16. LPIPS also uses the same 16-frame window. We report the checkpoint at each prescribed budget, without selecting the best checkpoint by validation score.

When rendering agent's pose as a skeleton, we use the following hyper-parameters: finger hue is set to $\phi_f = \phi_{\text{side}} + 24f$, with $\phi_L = 210$, $\phi_R = 0$, and $f = 0, \ldots, 4$ from thumb to little finger; left/right arm hues are $180/25$ degrees. The parameters $(\alpha_{\text{joint}}, \alpha_{\text{bone}}, \beta)$ are $(0.6, 0.9, 0.65)$ for hands and $(0.55, 0.75, 0.55)$ for arms.