

Action Response Is Not Factual Direction: A Prospective Audit of Robot Video World Models

Yinuo Chen

The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China

Abstract—Action-conditioned video models are often called controllable when changing the action changes the rollout, but response magnitude alone establishes neither movement toward the recorded future nor reliable sample ranking. We present ActionProbe, an artifact-audited protocol that separates branch response, signed factual-video advantage, one fixed composite repeatability contrast, and best-of-two sample-selection utility. On 200 command-enriched windows from 40 episodes and four conditioner-held-out tasks, decoded-RGB response–advantage correlations are positive in two fixed task–donor cells and negative in two. Across three donors and decoded-RGB, DINOv2 semantic/spatial, and RAFT motion endpoints, the fold-stratified raw-scale advantage interval includes zero at every endpoint, whereas normalized RGB and RAFT advantages are small and positive; a post-hoc fixed-task-weight sensitivity further changes which lower bounds clear zero. Donor-to-repeat response ratios range from 0.505 to 0.744, with every episode-bootstrap interval below one, conditional on a mixed-campaign design whose repeat branch also changes seed. A frozen generator-residual selector likewise fails its prospective gate: fold gains are +0.350 and −0.062, and the 180-window diagnostic gain is 0.144 [−0.140, 0.424]. Model-internal sensitivity should therefore not serve as factual-direction or sample-selection evidence without task-stratified, clustered, and prospective audit.

I. INTRODUCTION

Video world models are increasingly studied as data engines, predictive simulators, policy-evaluation tools, and action-selection surrogates for embodied agents [1], [2]. Before a generated rollout informs such decisions, one must know what its internally observable signals actually certify. A common controllability demonstration changes an action and shows that the generated video also changes. That test is neither a repeatability comparison nor a directional test: the change can be smaller than a repeated same-action rollout, move the wrong object, or drift farther from the observed future.

Offline robot data make this gap easy to overlook. For a recorded state, the dataset normally provides only one factual action and future. A zero or cross-task action therefore has no observed counterfactual future. We treat action replacement as a controlled stress intervention on the prediction and retain the recorded future as the sole factual target. This separates *response*, how much the generated video changes, from *factual directionality*, whether the change moves toward or away from that one recorded video. It does not evaluate correctness under the substituted action, whose counterfactual future is unobserved. Throughout, response is an operational

branch-separation statistic under nominally fixed sampling, not an isolated causal action effect.

The distinction is formal. With the future-frame MAE definitions in Sec. III-C, the reverse triangle inequality gives $|A_c| \leq R_c$. Response bounds the magnitude of a factual-trajectory advantage but supplies neither its sign nor a repeatability reference. We ask whether response exceeds one predeclared composite same-action contrast, whether its direction is stable across donors and representations, and whether a generator-native residual can rank two matched-action rollouts.

We answer these questions prospectively on WorldArena 2.0. An initial held-out fold is followed by a panel on two different evaluation tasks whose rule and synthesis were frozen before its outcomes were opened. The two folds use separately trained action-conditioned checkpoints, but remain one backbone family rather than an independent multi-model replication. Three donor instances per target, all drawn from that target’s one fixed donor task, preserve the same positive/negative RGB task pattern, but action-donor separation is only 50–74% of a fixed cross-campaign same-action contrast across four visual endpoints. Meanwhile, a training-selected generator self-score that looks useful on a 20-window pilot fails its predeclared primary test and does not recover on the new task fold.

Together, the intervention and selection audits ask whether two internal model signals—branch response and generator residual—support stronger factual-direction or sample-selection claims.

Our contributions are:

- an artifact-audited framework separating response, signed factual-video advantage, and one fixed repeatability contrast with episode-clustered inference;
- evidence that response–advantage relations change sign across fixed task–donor cells and that pooled direction depends on representation, scale, and task weighting; and
- a prospective falsification of a plausible generator-native best-of-two sample ranker, including why above-half accuracy can still yield negative expected gain.

II. RELATED WORK AND POSITIONING

a) *Robot video world models*: Action-conditioned visual prediction has a history in robot control and stochastic modeling [3], [4], [5]. WorldArena 2.0 broadens evaluation

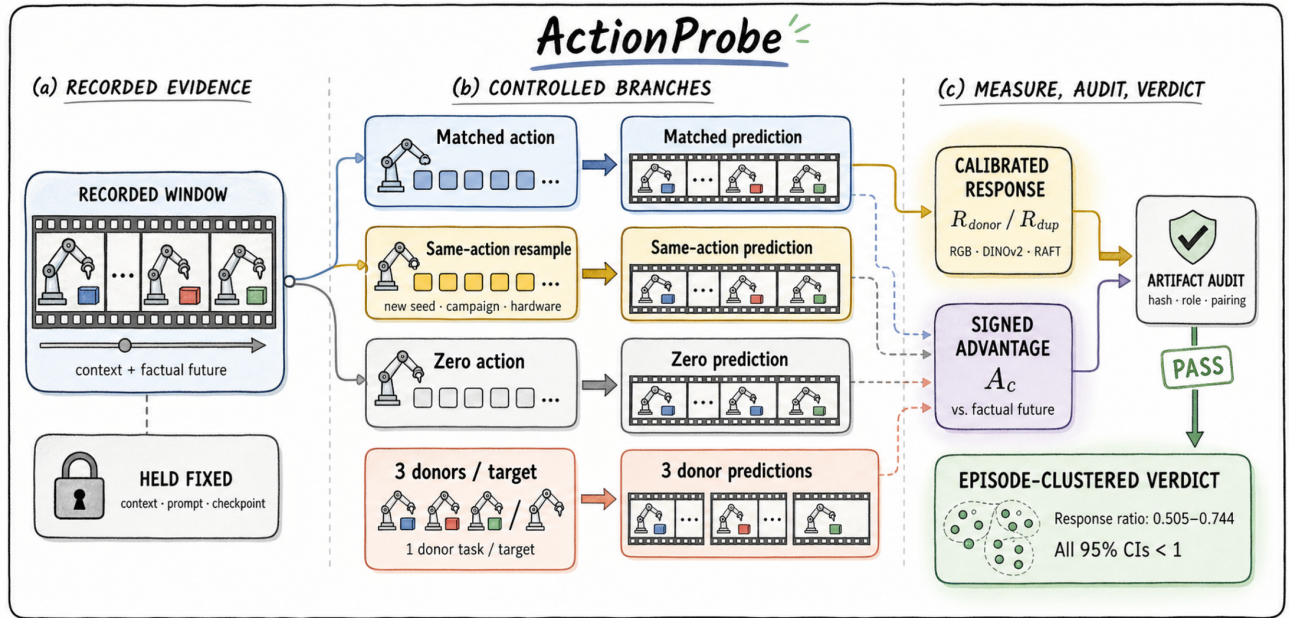


Fig. 1. **ActionProbe separates action response from factual direction.** One recorded context feeds matched, same-action-resampled, zero, and three cross-action donor branches per target. Donor 0 shares the original seed and campaign; donors 1–2 share the seed but not campaign/hardware, and the fixed repeat additionally changes seed. Response is compared in RGB, DINOv2, and RAFT spaces, whereas signed advantage compares donors with the sole factual future. Zero and donor actions remain stress interventions, not ground-truth counterfactual futures. The image’s “calibrated response” label means comparison with this one fixed composite reference only; the green “PASS” badge denotes artifact checks, not a scientific verdict. *AI-assisted illustration:* This figure was prepared using GPT Image 2 to visualize the protocol; the artwork does not constitute experimental evidence.

across modalities, functions, and platforms [1]; Cosmos-Predict2.5 provides open 2B and 14B video foundation models [2]. We do not propose a new backbone or report an official hidden-test score. We diagnose an action-conditioned 2B post-training instance using internal factual-video metrics. This position matters because perceptual quality, controllability, and usefulness as an action-selection surrogate are different evaluation targets. A visually plausible rollout may be command-sensitive yet still be an invalid surrogate for the recorded trajectory [6].

b) Action interventions and validity: CoCo and CD-LAM use same-state, zero-action, or transferred-action diagnostics [7], [8]. The admissibility ladder distinguishes action differentiation from validity [9]; CoWAM, WorldEval, and WPE audit selection or policy evaluation from generated rollouts [10], [11], [12]. WAV and SC3-Eval introduce learned forward-inverse or self-consistency checks [13], [14]; RoboWM-Bench and KineBench instead move toward grounded action recovery and simulator execution [15], [16]. We claim neither the invention of action intervention nor physical executability. Our narrower contribution is an auditable factual-GT implementation with same-noise pairing, an outcome-unopened evaluation panel, and episode-level uncertainty. Unlike simulator interventions, our substituted branches have no observed counterfactual target; unlike binary action-recovery tests, our signed quantity asks only whether a branch moves closer to the one factual video. Those restrictions make the estimand weaker but directly computable from offline robot data.

c) Generator self-evaluation: Best-of- N generation and model-native scoring are established. Most directly, Proprio scores flow residuals under controlled latent perturbations for video selection or refinement [17]. We transfer that principle to the frozen robot video generator as a baseline, then test whether its pilot benefit survives a predeclared evaluation-panel transfer and rerun checks. The contribution is therefore not a new residual formula: it is the prospective evidence that a plausible no-reference score can look useful on a small pilot while failing the outcome-unopened gates that would justify deployment.

III. EXPERIMENTAL PROTOCOL

A. Questions and Estimands

We organize the audit around three questions. *RQ1* asks whether cross-action branch separation exceeds a fixed cross-campaign same-action contrast. *RQ2* asks whether the matched branch is closer than intervened branches to the sole recorded future, and whether that direction persists across donors and representations. *RQ3* asks whether a no-reference generator residual can rank two matched-action samples before either is compared with that future. The questions share media but not estimands; we predeclare the episode as the replication unit for all three.

B. Data and Model

We pin the Apache-2.0 WorldArena 2.0 snapshot at revision `af1ac34d...`. It contains ten selected tasks with ten episodes per task. Four training tasks are shared: chip picking, coffee and water pouring, and table wiping. Fold 0

additionally trains on insertion and cucumber peeling and evaluates on box and shirt folding; fold 1 instead adds two table-cleaning tasks and evaluates on insertion and peeling. Thus the evaluation panels are disjoint and each is held out from its own conditioner, but fold 1 evaluation tasks occur in fold 0 training and residual-rule development. “Task shift” therefore means a separately trained fold 1 conditioner and unopened new evaluation panel, not method-development-unseen tasks.

Both folds use a Cosmos-Predict2.5-2B backbone with a 14D absolute-joint action conditioner. Arm angles are divided by π ; gripper opening is left in its native $[0, 1]$ scale or divided by 0.1 m, depending on the source schema. Each fold uses a 5,000-iteration all-legal-window checkpoint initialized from a 1,000-iteration balanced warm-up with 16 windows per training episode (batch 1, learning rate $2^{-14.5}$, weight decay 0.1). Each example contains one context frame and 12 future frames at 256×320 and 24 FPS. Generation uses the pinned source defaults: empty text prompt, 35 inference steps, and guidance 7.0. These values are bound by the hashed generation source; the raw media audits bind output roles and checkpoint paths but do not independently expose sampler arguments. The paired calls allocate about 8.18 GiB peak GPU memory per sample, so the complete audit is inference-heavy rather than a large-model training contribution.

a) *Evidence ladder*: Rule fit and development use 40 and 20 training-task windows. Fold 0 rank 0 then supplies a 20-window selector pilot; ranks 1–4 supply the outcome-unopened 80-window primary test. Those subsets share the same 20 episodes. The fold 0 ActionProbe diagnostic uses all 100 windows and therefore includes the opened rank 0 subset. Fold 1 contributes 100 panel-disjoint windows from 20 episodes whose outcomes were unopened when the transferred rule and synthesis were frozen. The nominal rerun revisits the same 20 pilot keys, while the 180-window pooled selector diagnostic excludes the pilot. “Unopened” always describes panel outcomes, not method-development-unseen tasks.

C. ActionProbe

For every recorded window, we generate a *matched* branch with its 12-step action chunk, a *zero* branch with an all-zero chunk, and a *cross* branch with a fixed action from the other task in the same fold (Fig. 1). The context frame, prompt, checkpoint, sampler configuration, and random seed are identical across branches. The cross branch is a stress intervention, never counterfactual ground truth. Zero is the literal all-zero normalized absolute-position command (zero arm angles and closed grippers), not a state-preserving no-op. For cross pairing, episodes are sorted by identifier and paired at equal episode rank; their five selected windows are paired at equal command-score rank. Each task therefore has one fixed donor task, not a balanced sample of admissible actions.

For intervention $c \in \{\text{zero}, \text{cross}\}$, define

$$R_c = \text{MAE}(\hat{v}_{\text{matched}}, \hat{v}_c), \quad (1)$$

$$A_c = \text{MAE}(v, \hat{v}_c) - \text{MAE}(v, \hat{v}_{\text{matched}}), \quad (2)$$

over future frames. Thus $R_c \geq 0$, while positive A_c means the recorded action is closer to the factual future v . MAE is averaged over decoded RGB values on $[0, 255]$, without semantic masking. All 400 matched–intervention comparisons satisfy $|A_c| \leq R_c$.

Each evaluation fold contains five non-overlapping windows per episode, selected by a command-only motion score: 100 windows in 20 episode clusters. This enrichment creates a high-action diagnostic panel, not an unbiased benchmark sample. Plans bind the panel order and fixed cross-task pairing before generation.

D. Frozen Same-Action and Multi-Donor Extension

We retain each original matched, zero, and cross output generated with seed 20260806, relabeling the cross branch as donor 0. Before opening extension outcomes, we freeze a same-action resample with seed 20260809 and two additional cross-task donors. Donors 1–2 come from distinct episodes of donor 0’s task and minimize absolute log-distance in the command-only motion score, with deterministic tie-breaking. They use seed 20260806 but, like the same-action output, were generated in the later A40 campaign; relative to the original matched branch they share a nominal seed but not campaign or hardware provenance. The same-action output additionally changes the diffusion seed, making it a fixed cross-campaign same-action contrast rather than a pure seed or kernel ablation. An independent audit verifies all 600 new videos, action hashes, roles, seeds, shapes, and frozen-plan identities.

For endpoint d and each window, let R_d^{dup} be matched-to-same-action distance, $\bar{R}_d^{\text{donor}}$ the mean distance to three donors, and $Q_d = \bar{R}_d^{\text{donor}} / R_d^{\text{dup}}$. Reported Q_d is the equal-episode mean of these window-level ratios, with a fold-stratified episode-bootstrap interval; it is not the ratio of the two aggregate distances reported later. We evaluate decoded-RGB MAE, DINOv2-Base CLS and aligned-patch cosine distance [18], and normalized distance between pooled RAFT-Small motion features [19]. Direction uses the analogous donor-mean advantage against the recorded future; its scale-free version divides each window by matched factual error. None of these features is a task-success or physical-correctness oracle. Every Q_d interval resamples episodes conditional on this fixed mixed-provenance design; it does not estimate variability across seeds, campaigns, hardware, or kernels. The fixed target–donor task mappings are box←shirt, shirt←box, insert←peel, and peel←insert.

E. Prospective Candidate Selection

We generate two matched-action candidates per window with seeds 20260806 and 20260807. Random-prior error is their mean factual error; oracle error is their minimum and is used only to measure available headroom after both outputs exist.

For the model-native baseline, candidate k is re-encoded as latent z_k after replacing its context frame by the exact observation. One Gaussian perturbation $\epsilon \sim \mathcal{N}(0, I)$ uses PyTorch seed 20260806 + 1000j for panel index j and is

shared by the candidate pair. At normalized flow times $\mathcal{T} = \{0.25, 0.50, 0.75\}$, the frozen Cosmos scheduler interpolates z_k and ϵ to $x_{k,\tau}$ and maps τ to its discrete timestep. Let $e_{k,\tau} = \hat{u}_\theta(x_{k,\tau}, \tau; a) - (\epsilon - z_k)$ be the matched-action denoising-velocity residual. The selected score is

$$S_k = \frac{1}{3} \sum_{\tau \in \mathcal{T}} Q(e_{k,\tau}; w_k),$$

$$Q(e; w) = \frac{\sum_{i,c} w_i e_{ic}^2}{C \sum_i w_i}, \quad (3)$$

$$w_{ki} = m_i [1 + 4 \text{clip}(d_{ki}/\bar{d}_k, 0, 10)].$$

where i indexes latent space–time cells, c indexes $C = 16$ channels, $m_i = 1$ only on future cells, and d_{ki} is the channel-mean $|z_{ki} - z_{k0}|$, and $\bar{d}_k$ averages d_{ki} over those cells (floored at 10^{-6} before division); the future mask always gives Q positive support. Lower S_k wins and a candidate-score tie selects the first seed. On fixed training-task splits of 40 fit and 20 development windows, six of 16 prespecified rules tied on gain and accuracy; the frozen deterministic tie-break (gain, accuracy, then lexicographically greatest rule name) selected `self_motion`. The top-ranked window from each fold 0 episode forms a 20-window first validation pilot; ranks 1–4 form an unopened 80-window primary test. These ranks do not overlap, but they come from the same 20 episodes, so the primary test is an outcome-unopened within-episode holdout rather than an episode-independent confirmation. The rule then transfers unchanged to all 100 windows of fold 1. The two-fold pooled diagnostic therefore contains 180 windows: it excludes the 20 fold 0 pilot windows. Pooled results weight the 40 episode means equally; the corresponding raw-window mean gain is 0.121.

F. Integrity and Statistics

The evidence package records dataset revision and splits in a protocol manifest; raw audits bind checkpoints, action sources, panel plans, and decoded outputs, while hashed generation source defines the held-fixed branch-call contract; every declared claim source carries a SHA-256 hash. The anonymous supplement contains the complete revision and splits, launch and scoring source, raw per-window outputs, and manifests. Independent audits re-decode videos and recompute labels; selector features do not access the factual future.

The episode, not the window, is the independent resampling unit. We use 10,000 percentile nonparametric episode-bootstrap draws with deterministic streams derived from base seed 20260806 for the core analyses; R041 robustness uses streams derived from 20260809, and post-hoc R042 records seeds 20260810–20260825. Primary pooled analyses resample within fold, and Spearman’s ρ_s is recomputed within each draw. Because within-fold resampling also lets the observed task proportions fluctuate, a post-hoc sensitivity instead resamples within each fixed task and gives the four task means equal weight. It never replaces a predeclared gate. Task intervals and the observed maximum-minus-minimum task span are descriptive. Task labels, donors, scene statistics, and error scales are structural, so no population-of-tasks claim is made.

The fold 0 selector gate jointly required: positive mean gain, a cluster-CI lower bound above zero, point-estimate accuracy above 0.5, and lower MAE than the training-selected fixed seed B. Accuracy itself is also reported with a cluster interval. Fold 1 and pooled analyses are diagnostic transfer and synthesis; they cannot change the fold 0 primary verdict.

For transparency, let g_{efj} denote random-prior MAE minus selected MAE for window j in episode e and fold f . The reported pooled selector estimator is

$$\hat{G} = \frac{1}{40} \sum_{f=1}^2 \sum_{e=1}^{20} \left(\frac{1}{m_{ef}} \sum_{j=1}^{m_{ef}} g_{efj} \right), \quad (4)$$

where $m_{ef} = 4$ for fold 0 after excluding the pilot rank and $m_{ef} = 5$ for fold 1. Action advantages use analogous episode means: task summaries are stratified by task and the pooled interval by fold. The heterogeneity statistic is the maximum minus minimum of the four task means, not a test on the four response–advantage correlations.

G. Audit Surface and Reproduction Boundary

Records freeze keys, donors, seeds, rules, media roles, windows, and checkpoints; scripts recompute endpoints and intervals; and a claim manifest maps principal numbers to hashed artifacts. The anonymous archive bundles plans, per-window tables, selected scripts, manifests, and Fig. 5’s eight MP4s, but omits other videos, tensors, weights, and checkpoints. It supports bundled recomputation, not end-to-end regeneration. The first campaign did not freeze every kernel, driver, GPU, or container digest, so reruns never replace primary outputs; two near-zero selector margins flip despite nearly stable error labels.

IV. RESULTS

Response fails two stronger interpretations in these data: its factual direction changes across fixed task–donor cells, and its magnitude is below one fixed composite same-action contrast. Neither result establishes absence of action conditioning. We then ask whether a generator-internal residual nevertheless supports prospective sample selection.

A. Response Does Not Establish Factual Direction

The pooled matched-versus-cross advantage is 0.253 $[-0.023, 0.581]$, so these data do not establish a stable positive aggregate benefit. Aggregation also hides structured sign changes. Across the four fixed task–donor cells, cross-action means span -0.241 to $+1.162$; the descriptive maximum-minus-minimum span is 1.403. Cell-to-cell variation also appears against zero actions (Table I). These contrasts are conditional on the four tasks, donor identities, and predefined pairings, not population-level evidence over robot tasks.

Response itself is not a directional certificate. In the frozen two-fold synthesis, clustered response–advantage correlations are 0.516 with interval $[0.234, 0.745]$ for box folding and -0.418 with interval $[-0.674, -0.085]$ for shirt folding. In fold 1, they are -0.450 with interval $[-0.667, -0.200]$ for insertion and 0.482 with interval $[0.233, 0.675]$ for cucumber

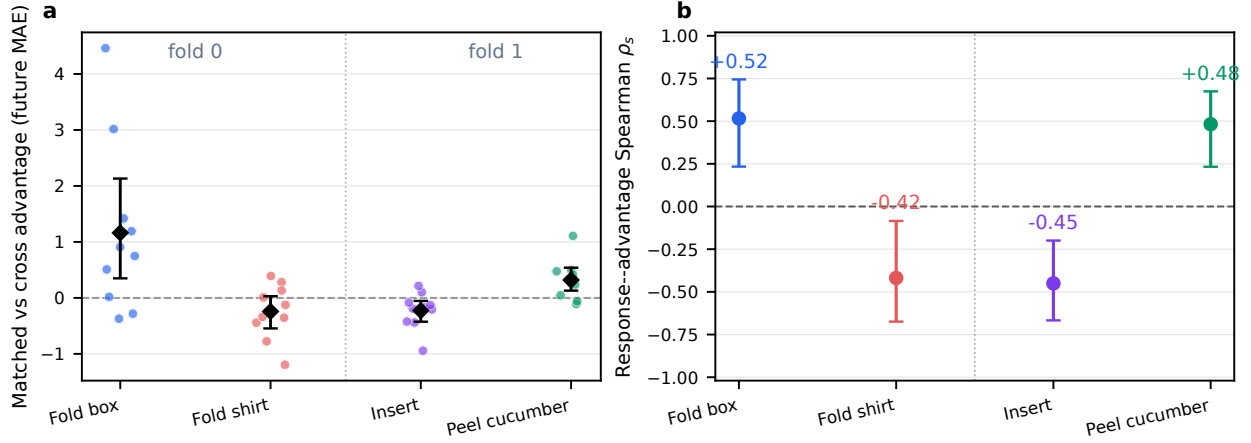


Fig. 2. **Observed response-advantage correlations include both signs across fixed task-intervention cells.** **a**, Episode means (five frozen windows each), task means (diamonds), and 95% episode-bootstrap intervals for matched-versus-cross future-MAE advantage. **b**, window-level Spearman correlations with episode-cluster bootstrap intervals. Fold 0 contains the initial panel; fold 1 is an outcome-unopened evaluation panel with a separately trained conditioner. Together: $n = 200$ windows, 40 episodes, four tasks. Conditional on these tasks and fixed pairings, task effects and correlation intervals are descriptive rather than a population-of-tasks inference.

TABLE I

SIGNED MATCHED-ACTION ADVANTAGE UNDER FIXED STRESS INTERVENTIONS. POSITIVE VALUES FAVOR THE RECORDED ACTION IN FUTURE MAE. INTERVALS RESAMPLE EPISODES; TASK-LEVEL CROSS INTERVALS USE THE FROZEN TWO-FOLD SYNTHESIS. THE POOLED CROSS ROW IS STRATIFIED BY FOLD. ZERO-ACTION RESULTS ARE RETAINED AS A ROBUSTNESS CONTRAST.

Fold	Panel	n	vs. zero: mean [95% CI]	vs. cross: mean [95% CI]
0	Overall	100	-0.261[-0.560, 0.082]	+0.460[-0.036, 1.073]
0	Fold box	50	+0.105[-0.413, 0.653]	+1.162[0.350, 2.132]
0	Fold shirt	50	-0.627[-0.813, -0.435]	-0.241[-0.545, 0.030]
1	Overall	100	-0.106[-0.445, 0.155]	+0.047[-0.135, 0.230]
1	Insert	50	-0.480[-1.017, -0.073]	-0.227[-0.425, -0.054]
1	Peel cucumber	50	+0.268[0.135, 0.406]	+0.320[0.130, 0.540]
0-1	Pooled diagnostic	200	—	+0.253[-0.023, 0.581]

peeling (Fig. 2b). Thus larger output changes align with better factual-video predictions on two tasks and worse predictions on two, despite the same metric definitions and held-fixed controls.

B. The Fixed Same-Action Contrast Exceeds Donor Response

The multi-donor extension changes the interpretation of nominal response (Fig. 3). Episode-mean Q_d is 0.505 [0.478, 0.530] for decoded RGB, 0.628 [0.583, 0.682] for DINOv2 CLS, 0.672 [0.648, 0.697] for aligned patches, and 0.744 [0.693, 0.801] for RAFT motion. Every interval lies below one: across all four endpoints and conditional on this fixed comparison, the three-donor separation under a shared nominal seed but mixed campaign/hardware provenance is smaller than the later same-action resampling/campaign separation. This does not prove that the action conditioner has no effect; it shows only that uncalibrated branch separation is not larger than this frozen operational contrast.

The corresponding absolute episode means (same-action repeat versus three-donor response) are 8.859 versus 3.722 for RGB, 0.150 versus 0.084 for DINOv2 CLS, 0.192 versus 0.121 for DINOv2 patches, and 0.410 versus 0.227 for RAFT. These endpoint-specific scales show that $Q_d < 1$ reflects

substantial repeat drift rather than a dimensionless absence of response.

The donor-dependent RGB pattern is nevertheless stable. All three donors give positive task-mean advantage for box folding and cucumber peeling and negative advantage for shirt folding and insertion. Pairwise donor rank correlations are 0.909–0.933 for response and 0.724–0.743 for advantage. Yet pooled donor-mean raw-scale advantage intervals include zero for decoded RGB, both DINOv2 endpoints, and RAFT. After per-window normalization by matched factual error, RGB and RAFT have small positive means, 2.54% [0.62, 4.58] and 2.82% [0.08, 6.20], whereas both DINOv2 intervals cross zero. Across these three selected donors within the fixed donor task, changing donor episode/window identity does not reverse the original RGB task signs, but semantic and motion representations do not provide uniform directional confirmation.

The fixed-task-weight sensitivity exposes the cost of treating the observed task mix as exchangeable. It leaves every Q_d upper bound below one (0.528, 0.675, 0.693, and 0.794), but gives decoded-RGB raw advantage a positive lower bound, 0.254 [0.034, 0.493], while raw DINOv2 CLS, DINOv2 patch,

TABLE II

FROZEN SELECTOR GAINS BY FOLD AND METRIC. MAE GAIN IS RANDOM MINUS SELECTED; PSNR/SSIM GAINS ARE SELECTED MINUS RANDOM. INTERVALS RESAMPLE EPISODES; EACH JOINT GATE IS NO-GO.

Subset	Metric gain [95% CI]	LB > 0
Fold 0 unseen 80	MAE +0.350[−0.187, 0.883]	fail
	PSNR +0.113[−0.104, 0.326]	fail
	SSIM +0.00632[0.00246, 0.01010]	pass
	Joint metric gate	NO-GO
Fold 1 all 100	MAE −0.062[−0.239, 0.101]	fail
	PSNR +0.091[−0.108, 0.303]	fail
	SSIM +0.00126[−0.00037, 0.00286]	fail
	Joint metric gate	NO-GO

and RAFT intervals still cross zero; all four normalized lower bounds become positive. Exact pooled directional inference is therefore weighting- and scale-dependent, reinforcing task- and representation-stratified reporting rather than a universal claim.

C. Prospective Self-Ranking Audit

On the 20-window fold 0 pilot, the two seeds provide a 9.59% oracle-over-random MAE gap. The training-selected residual rule appears to recover 57.0% of that gap, with gain 1.076 [0.108, 2.070]. This first validation result is hypothesis-generating: it motivated carrying the already training-selected rule into the frozen primary and fold 1 evaluation panel.

The prospectively frozen 80-window primary result is weaker. Mean gain remains positive at 0.350, but its episode-cluster interval [−0.187, 0.883] crosses zero. The four gate criteria are, respectively: positive mean (pass), CI lower bound above zero (fail), point accuracy $0.5625 > 0.5$ (pass), and selected MAE 16.089 below fixed-seed-B MAE 16.126 (pass). The accuracy cluster interval [0.438, 0.688] crosses 0.5, so the point criterion is not evidence of above-chance accuracy. The joint primary verdict is therefore NO-GO.

Transfer to the prospectively frozen new evaluation-task fold does not rescue the score: its all-window MAE gain is −0.062[−0.239, 0.101]. The fold-1-minus-fold-0 gain contrast is −0.412[−0.968, 0.158], which is also inconclusive. The equal-episode-weighted diagnostic pooled gain over 180 windows and 40 episode clusters is 0.144 [−0.140, 0.424]; this pool excludes the fold 0 pilot and is not another sequential confirmation stage. Metric changes do not rescue the method: fold 0 SSIM alone passes its lower-bound criterion, whereas PSNR fails and both fold 1 metric criteria fail (Table II).

A post-hoc fixed-task-weight sensitivity preserves the same no-go inference: fold 0, fold 1, and pooled MAE-gain intervals are respectively [−0.129, 0.787], [−0.240, 0.098], and [−0.104, 0.380].

The candidate pool contains meaningful but fold-dependent oracle headroom. On fold 0 unseen ranks, random and oracle MAE are 16.438 and 14.882, so the selector recovers 22.5% of the available two-sample gap. On fold 1 they are 6.399 and 5.982, but the negative selector gain corresponds to recovering −14.9% of that smaller gap. Therefore the transfer failure cannot be explained by identical candidates alone: an oracle

can still improve both folds, while the frozen no-reference score orders them unreliably.

Point accuracy and expected gain need not agree because mistakes have unequal cost. In the post-hoc magnitude decomposition, fold 1 chooses the lower-MAE candidate on 58/100 windows, but those correct choices improve MAE by only 0.306 on average; its 42 mistakes cost 0.571 on average, producing the negative −0.062 mean. Fold 0 instead has 45/80 correct choices with mean gain 1.694 and 35 mistakes with mean gain −1.379, producing +0.350. This decomposition explains the no-go outcomes but is not an additional predeclared gate. A selector intended for policy screening therefore needs magnitude-sensitive risk, not only more-than-half pairwise accuracy.

This mismatch has a simple decision-theoretic form. If C denotes choosing the lower-error candidate, $\pi = \Pr(C)$, and μ_C, μ_{-C} are the conditional gains, then

$$G = \pi\mu_C + (1 - \pi)\mu_{-C}. \quad (5)$$

Positive accuracy is insufficient whenever the cost-weighted second term dominates. Given the observed wrong-choice costs, fold 0 needs conditional correct-choice gain above 1.072 and attains 1.694; fold 1 needs above 0.413 but attains only 0.306. The failure is therefore not merely “too few” correct choices: the score is right on smaller-gap pairs and wrong on larger-gap pairs. For best-of-two sample ranking, a selector should report expected gain or regret jointly with accuracy and should calibrate any score-margin abstention rule on separate data. We do not fit such a rule here, because doing so after opening these panels would create a new, result-dependent method.

a) Development landscape and ablation: The selected score is not uniquely favored even on its development set. Pure future residual, motion-weighted residual, and four weakly weighted hybrid rules make exactly the same choices, yielding 0.65 accuracy and 0.511 MAE gain on 20 development windows. In contrast, pure action-contrast rules have gains between −0.376 and −0.574 there; larger hybrid coefficients progressively erase the apparent benefit. This small opened set only selects a rule and is not inferential evidence. On the frozen fold 0 unseen ranks, replacing `self_motion` by the unweighted `self_future` changes selected MAE by −0.0127[−0.0373, 0.0003] (motion minus future improvement). On fold 1 the change is −0.0167[−0.0455, −0.0006]. Motion weighting therefore does not explain the pilot success and slightly worsens the transferred selection.

The same-example rerun diagnoses implementation-level reproducibility. With the same keys, checkpoint, action chunks, sampler seeds, guidance, and residual protocol, both candidates are bitwise identical for 12/20 windows; for the other eight, prior-versus-rerun future-video MAE is nonzero. Across all 40 candidate outputs it averages 0.304 and reaches 1.422, which is not negligible beside the fold 1 task-mean branch separations of 1.748–1.895. Candidate factual-error labels nevertheless change by only 0.0024 MAE on average (maximum 0.0228). Two near-tied score margins cross zero and choices flip. Gain shrinks from 1.076 to 0.624, with

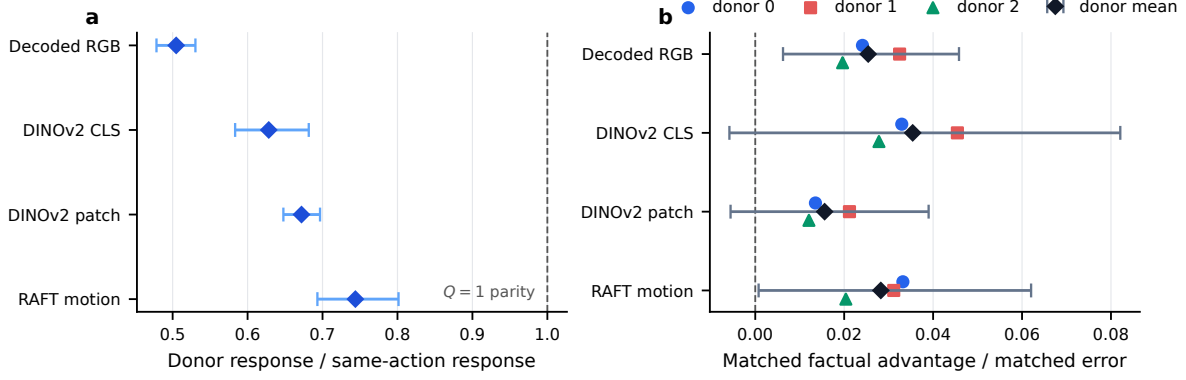


Fig. 3. **Cross-action response is smaller than a fixed cross-campaign same-action contrast on every endpoint.** **a**, Episode-mean response ratio Q_d with 95% stratified episode-bootstrap intervals; values below one favor the fixed same-action contrast. **b**, donor-specific and donor-mean factual advantage normalized by matched factual error. RGB is decoded MAE, DINOv2 uses CLS and aligned-patch cosine distances, and RAFT uses pooled motion-feature distance. Donors 1–2 and the repeat were generated in a later A40 campaign; the repeat also changes seed. Thus Q_d is one mixed-provenance operational contrast, not an estimate of kernel nondeterminism or a causal action effect.

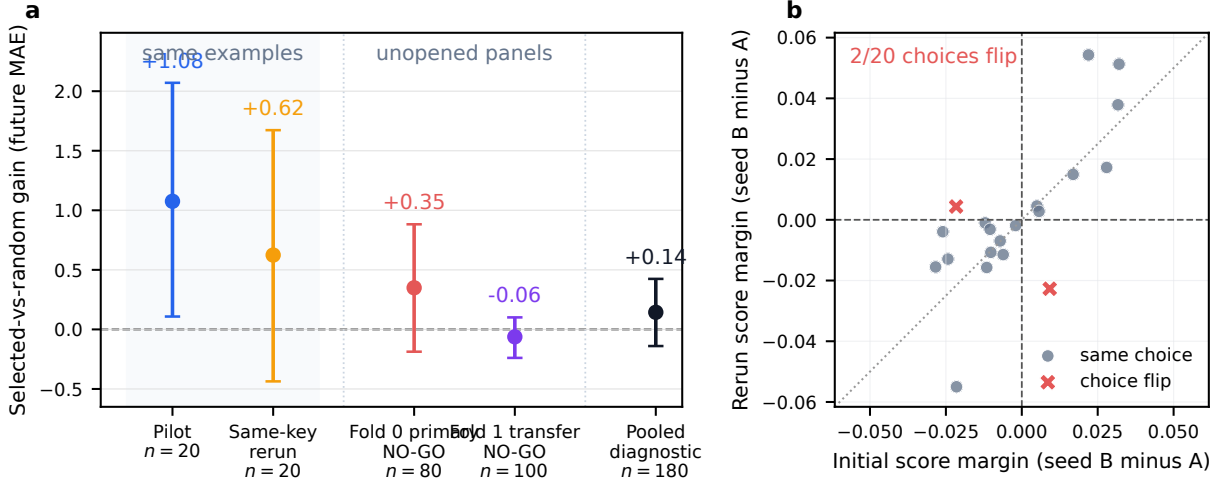


Fig. 4. **Pilot evidence does not survive prospective and rerun checks.** **a**, selected-versus-random MAE gain. The first two points use the same 20 fold 0 examples; fold 0 frozen and the fold 1 evaluation panel are outcome-unopened tests. Pilot, rerun, and prospective-fold intervals come from their frozen single-stage artifacts; the visually separated equal-episode-weighted 180-window pooled diagnostic uses the frozen two-fold synthesis. The pool excludes pilot rank 0. **b**, a nominal same-key regeneration flips 2/20 choices near zero score margin despite nearly unchanged candidate factual-error labels. Output-video instability is quantified in the text. Episode is the uncertainty unit.

cluster interval $[-0.436, 1.673]$. Deterministic GPU kernels and full hardware/software identity were not bound, so this audit detects but does not attribute the nondeterminism. It is not a new panel, and without a within-campaign duplicate same-action baseline it prevents a clean action-causal interpretation of R_c .

The grid in Fig. 5 is not used to score physical correctness. It makes the audited media inspectable while the estimand remains distance to the sole recorded future, and its mechanical selector prevents choosing visually dramatic cases after seeing outcomes.

V. DISCUSSION, ROBUSTNESS, AND LIMITATIONS

a) Response is capacity, not direction: The inequality $|A_c| \leq R_c$ explains why zero branch separation precludes a large factual advantage, but also why large response alone is

ambiguous. The differing taskwise signs make this ambiguity operational: mean-only controllability summaries can hide qualitatively different relationships between action sensitivity and the factual video.

b) Metric and rule checks: Changing the selector endpoint does not jointly rescue it: both folds fail the PSNR/SSIM gate, although fold 0 SSIM alone has a positive lower bound. The training-side rule is itself weakly identified: six of 16 prespecified rules tie on development gain and accuracy, so the deterministic name tie-break selects `self_motion`. This is valid for a frozen test, but it argues against treating the chosen residual as a uniquely supported mechanism.

c) Evaluation implication: For open-loop robot video, we recommend reporting four items together: (i) paired nominal-sampling branch separation, (ii) signed advantage against the sole factual future, (iii) task-stratified episode-

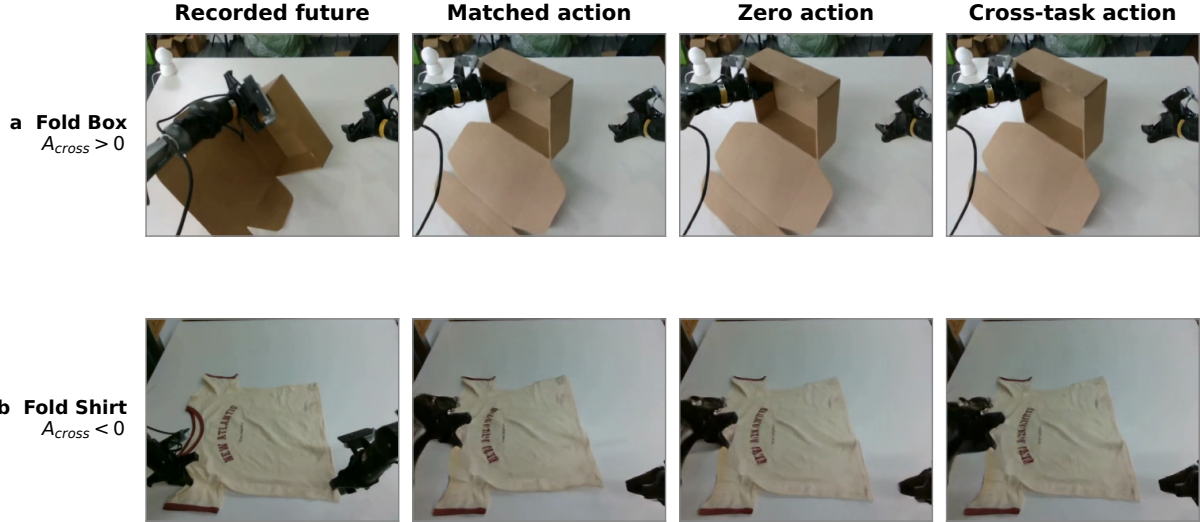


Fig. 5. **Mechanically selected fold 0 examples illustrate why visual response is not a sign test.** Columns show final frames from audited 13-frame videos. Per task, the selected episode has mean cross advantage closest to the task episode-mean; within it, the window is closest to that episode mean. No appearance-based selection is used. The box case has positive A_{cross} , whereas the shirt case has negative A_{cross} . These two cells are illustrations, not semantic or counterfactual labels; temporal videos remain in the artifact.

cluster uncertainty, and (iv) prospective stability for any learned or generator-native selection score. A no-reference selector should also expose score margins and a nominal rerun, because small numerical changes can alter near-tied decisions.

d) A practical audit sequence: Freeze windows and donors before outcomes; compare paired response with multiple same-action seeds (here, one fixed cross-campaign contrast); then report signed A_c by donor, representation, task, and episode. For video ranking, freeze the score on training tasks and test expected gain, rerun exact keys, and finally open a panel-disjoint fold without revising the rule. This ordering keeps response, an opened pilot, or a metric-specific pass from being mistaken for prospective sample-ranking evidence.

e) What remains unidentified: The four cells show that one aggregate response statistic mixes positive and negative factual associations; they do not identify why. Baseline prediction difficulty, visual texture, donor-task support, action magnitude, and true task dynamics vary together. Our three donors vary episodes and windows but not the donor task. Randomizing admissible donor tasks, adding grounded object-state readouts, and repeating across independently designed world-model families are needed to attribute the pattern rather than merely diagnose it.

f) Limitations: The study has four conditioner-held-out tasks, 40 episodes, enriched windows, one backbone family, and two separately trained checkpoints—not a multi-model replication. RGB, DINOv2, and RAFT are factual-video diagnostics, not grounded success, physical correctness, or the full WorldArena score [20], [21]. Zero need not be a no-op; one possibly out-of-support donor task per target confounds task, donor, and intervention magnitude; and cross actions lack counterfactual truth. Fold 0 pilot and primary ranks

share episodes. The mixed-campaign same-action control does not isolate a nondeterminism source; a 20-key same-seed rerun has mean/max MAE 0.304/1.422 and eight non-bitwise-identical pairs, so R_c is branch separation, not an isolated action effect. The open-loop selector has only $N = 2$ candidates and does not establish policy value or executability; 40 clusters limit precision.

Fold 1 is panel-disjoint and outcome-unopened with a separate conditioner, but its tasks occurred during fold 0 development. It is neither method-development-unseen nor an independent replication. Full bitwise reproduction would also require a frozen driver, kernel, hardware, and container stack.

VI. CONCLUSION

Across four visual endpoints, three-donor action response is smaller than a fixed composite same-action contrast, while decoded-RGB response–advantage relations have opposite signs across the four fixed task–donor cells. Pooled directional inference changes with scale and task weighting, and a transferred residual that appears useful on a small pilot fails both outcome-unopened panels. Operational response, factual direction, repeatability references, and generator sample-ranking performance should therefore be reported separately with task-stratified episode uncertainty and prospective gates within the tested four-task, one-backbone scope ($N = 2$ for selection).

REFERENCES

- [1] Y. Shang, Y. Tang, Y. Ma, Z. Li *et al.*, “WorldArena 2.0: Extending embodied world model benchmarking on modality, functionality and platform,” *arXiv preprint arXiv:2605.17912*, 2026.
- [2] NVIDIA, A. Ali, J. Bai, M. Bala *et al.*, “World simulation with video foundation models for physical AI,” *arXiv preprint arXiv:2511.00062*, 2025.

- [3] C. Finn and S. Levine, "Deep visual foresight for planning robot motion," in *Proc. IEEE Int. Conf. Robotics and Automation*, 2017.
- [4] F. Ebert, C. Finn, A. X. Lee, and S. Levine, "Self-supervised visual planning with temporal skip connections," in *Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 78, 2017, pp. 344–356.
- [5] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine, "Stochastic variational video prediction," in *International Conference on Learning Representations*, 2018.
- [6] H. Li, J. Wang, Z. Mei, A. Majumdar, J. Chen, and B. Zhu, "RoboTrustBench: Benchmarking the trustworthiness of video world models for robotic manipulation," *arXiv preprint arXiv:2606.01600*, 2026.
- [7] Y. Shi, Z. Chu, J. Wei, J. Hao, J. Liu, and J. Fu, "Overcoming statistical bias in action-controllable world models," *arXiv preprint arXiv:2608.04653*, 2026.
- [8] Y. Wei, K. Zhou, L. Mao, Z. Zhang, Z. Xu, Z. Xi, S. Liang, R. Han, Y. Yan, X. Wang, F. Feng, and B. Huang, "Causally debiased latent action model for embodied action conditioned world models," *arXiv preprint arXiv:2607.09185*, 2026.
- [9] C. Oefinger, F. R. Schäfer, K. Moller, M. Piccinini, and J. Betz, "Validate the dream before you trust its verdict: Admissibility for world-model simulators," *arXiv preprint arXiv:2607.07196*, 2026.
- [10] S. Liu, Q. Wen, S. Hao, Q. Luo, C. Zhang, F. You, C. Wu, and N. Su, "CoWAM: Coordination contracts for selective policy intervention with WAMs," *arXiv preprint arXiv:2608.02578*, 2026.
- [11] Y. Li, Y. Zhu, J. Wen, C. Shen, and Y. Xu, "WorldEval: World model as real-world robot policies evaluator," *arXiv preprint arXiv:2505.19017*, 2025.
- [12] J. H. Quevedo, A. K. Sharma, Y. Sun, V. Suryavanshi, P. Liang, and S. Yang, "WorldGym: World model as an environment for policy evaluation," in *International Conference on Learning Representations*, 2026. [Online]. Available: <https://openreview.net/forum?id=hidBHy1CAw>
- [13] Y. Liu, F. Feng, L. Kong, W. Lu, J. Tang, K. Zhang, K. Murphy, C. Finn, and Y. Du, "World action verifier: Self-improving world models via forward-inverse asymmetry," *arXiv preprint arXiv:2604.01985*, 2026.
- [14] W.-C. Tseng, G. Hussein, Y. Dong, A. Z. Ren, L. X. Shi, X. Wang, S. Levine, Z. Li, J. Gu, F. Shkurti, M.-Y. Liu, and Q. Vuong, "SC3-Eval: Evaluating robot foundation models via self-consistent video generation," *arXiv preprint arXiv:2606.18610*, 2026.
- [15] F. Jiang, Y. Chen, K. Xu, Y. Liu, H. Wang, Z. Shen, J. Lu, S. Huang, Y. Wang, C. Xie, and R. Wu, "RoboWM-Bench: A benchmark for evaluating world models in robotic manipulation," *arXiv preprint arXiv:2604.19092*, 2026.
- [16] Z. Liu, Z. Zhu, Y. Zhang, C. Fan, C. Bai, and X. Li, "KineBench: Benchmarking embodied world models via IDM-free kinematic grounding," *arXiv preprint arXiv:2607.19876*, 2026.
- [17] M. Hassan, K. Messaoud, W. Li, and A. Alahi, "Proprio: Latent self-scoring and inference-time refinement for physically plausible video generation," *arXiv preprint arXiv:2605.28230*, 2026.
- [18] M. Oquab, T. Darcet, T. Moutakanni *et al.*, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024.
- [19] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *European Conference on Computer Vision*, 2020.
- [20] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [21] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.