WHEN DO SPATIAL COMMANDS STEER SPORTS VIDEO GENERATION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Video diffusion models can generate realistic futures from short input videos, yet they remain unreliable at following instructions about how objects should move. Despite rapid progress in controllable video generation, it remains unclear whether these instructions actually guide the generated motion or whether models mainly extrapolate from the input video. We begin by asking how much a generated object's motion changes when its instructed path changes. To separate instruction following from general future prediction, we consider two complementary tests: (1) response to intervention, where two generations share the same input video and randomness but receive different object paths, and (2) future similarity, where each generation is compared with the recorded continuation. Broadcast sports provide a natural testbed for this distinction because the movement of a ball or shuttlecock is central to what happens next. Using this setting, we introduce **SportWorldBench**, a controlled framework for evaluating object-motion control in video generation. Across changes to path labels, how paths are provided, model training, video resolution, and random samples, we find no reliable response to changed paths. Moreover, at higher resolution the models resemble the recorded future more closely than simply repeating the last input frame, yet this resemblance is nearly identical whether the instructed path matches the recording or is deliberately altered. These findings show that generating a plausible future does not imply following object-motion instructions and motivate methods that make such controls reliably influence generation.

1 INTRODUCTION

Video diffusion models have made striking progress in visual realism and temporal coherence (Yang et al., 2025; Jin et al., 2025; Chen et al., 2023), yet they remain unreliable when asked to follow precise object-motion controls. While representative video generators achieve strong subject and background consistency on VBench (Huang et al., 2024), trajectory-controlled systems remain substantially less reliable at aligning object motion with instructed paths (Li et al., 2025b). These shortcomings reveal a critical gap between preserving a video's appearance over time and making a generated object follow an instructed path. More fundamentally, proximity to the instructed path does not establish that the instruction caused the motion, because the input video may already imply the same path.

Previous works have studied whether video diffusion models can follow user-provided object trajectories (Shi et al., 2024; Wu et al., 2024; Zhang et al., 2025), or generate videos with strong visual and physical consistency (Huang et al., 2024; Li et al., 2025a), often through benchmarks that evaluate each output independently. Such studies are valuable for understanding the abilities and limitations of controllable video generation. However, most of these efforts do not investigate whether changing an object's instructed path actually changes that object's generated motion.

There are multiple possible reasons why object-motion control may fail in video diffusion models, such as how trajectories are represented and incorporated into the model (Wu et al., 2024; Zhang et al., 2025; Li et al., 2025b). In particular, we focus on whether the instructed path influences generated motion. We ask the following central question: *How reliably do video diffusion models respond to a different object path?* To separate command response from general future prediction, we pose two research questions.

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.

An output can lie near an instructed path because the input video already implies similar motion. We therefore formulate **(RQ1) Response to intervention:** Given the same input and randomness, does changing only the instructed object path change the generated object's motion? How does this response vary with the object-path annotations used for training and the way the path enters the model? We evaluate paired outputs, validate the test on real-video edits with known translations, and bound missing observations.

A model may also generate a plausible continuation while ignoring the instructed path. We therefore formulate **(RQ2) Future similarity:** Does a generated continuation resemble the recorded future more than a repeated-last-frame baseline, and does this similarity change when the instructed path changes? Because the recording is only one plausible future, this comparison measures prediction rather than adherence.

Broadcast sports provide a natural setting because the movement of a ball or shuttlecock is central to what happens next. These objects can also be small, fast, blurred, or hidden behind players (Tarashima et al., 2023; Gossard et al., 2026). We therefore introduce **SportWorldBench**, a controlled framework built from broadcast videos of soccer, basketball, badminton, and table tennis. For each window, paired generations differ only by a fixed horizontal shift of the object path, and a detector that never sees either path measures the resulting motion.

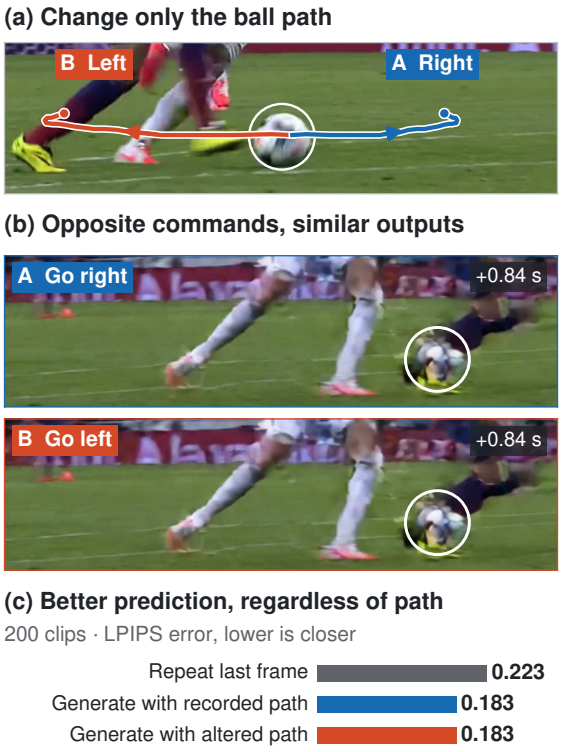

Figure 1: **Prediction is not path control.** (a) Opposite commands share the same input. (b) Paired 2B outputs at 0.84 s place the ball (circled) in nearly the same location, with model and sampling noise fixed. (c) A separate 200-clip benchmark shows 18% lower LPIPS than frame repetition under either recorded or shifted paths. The close-up itself does not improve LPIPS (0.448 versus 0.427 for repetition). See Appendix G for provenance.

For RQ1, we cross heuristic versus sport-specific training paths with flattened-token versus explicit causal-grid alignment. This separates the effect of path annotations from how paths enter the model. The adapters follow residual spatial-conditioning and LoRA designs (Zhang et al., 2023; Hu et al., 2022). We further vary adaptation recipe, video resolution, sampling seed, and object-channel presence. Differences between the two starting checkpoints extend beyond model size, so their comparison does not isolate capacity. Finally, frozen-model tests on animals and road vehicles examine whether the findings extend beyond sports.

Our findings are summarized as follows:

- **Response to intervention.** The test reliably detects known translations in edited real videos. In generated videos, however, none of the evaluated settings supports a reliable response to the changed path. This conclusion holds under worst-case treatment of missing observations.
- **Robustness across conditions.** Failure to respond persists across changes to object annotations, control alignment, adaptation recipe, resolution, random sample, and the presence of object-control channels. None of these changes consistently restores response.
- **Separation of future prediction and motion control.** At native resolution in sports, the models resemble the recorded future more closely than a repeated last frame. This advantage is nearly identical for matched and shifted paths, showing that future prediction does not imply path control.

2 Related Work

Video diffusion models. Diffusion models extended image architectures to coherent unconditional and text-conditioned videos (Ho et al., 2022). Latent models improved resolution through large-scale pretraining (Blattmann et al., 2023), while recent systems advanced long-video prediction, temporal coherence, and efficiency (Chen et al., 2023; Yang et al., 2025; Jin et al., 2025). World foundation models predict future observations from visual context and actions (NVIDIA, 2025a;b). These advances improve plausible generation without establishing whether spatial instructions determine the motion.

Object-motion control in video generation. Recent methods control video through trajectories, masks, entity features, or motion representations. DragNUWA combines text, an input image, and a trajectory, while MotionCtrl separates object and camera motion (Yin et al., 2023; Wang et al., 2024). Motion-I2V converts sparse trajectories into motion fields, DragAnything conditions on entity representations, and PEEKABOO uses spatiotemporal masks (Shi et al., 2024; Wu et al., 2024; Jain et al., 2024). Tora incorporates trajectories into a diffusion transformer, and MagicMotion studies dense-to-sparse guidance with trajectory-alignment measures (Zhang et al., 2025; Li et al., 2025b). Scoring each output against its control cannot distinguish instruction response from motion already implied by context. We instead compare paired generations with identical context, model, and sampling noise but different paths.

Evaluation and controlled diagnosis. Video metrics summarize distributional quality or perceptual similarity (Unterthiner et al., 2018; Zhang et al., 2018), while broader benchmarks assess temporal consistency, motion, text alignment, and physical plausibility (Liu et al., 2024; Huang et al., 2024; Li et al., 2025a). These properties do not identify which conditioning signal produced an object's motion. Following controlled diffusion studies that vary one source of evidence at a time (Jeong et al., 2026), we change only the object path and measure response separately from future similarity.

Small-object tracking in sports video. Balls and shuttlecocks can be small, blurred, occluded, and observed under camera motion. SoccerNet established large-scale broadcast-video evaluation (Giancola et al., 2018), WASB develops a shared tracking baseline across ball sports (Tarashima et al., 2023), and BlurBall addresses severe table-tennis motion blur (Gossard et al., 2026). We use tracking to observe a generator's response rather than as the target task. Known real-video transformations calibrate the measurement, while coverage and bounds expose missing observations.

3 SportWorldBench: Diagnosing Object-Motion Control

SportWorldBench distinguishes two abilities that are easily conflated in conditional video generation: producing a plausible continuation and responding to an instructed object path. Each item contains a real broadcast-video context, spatial controls, and a recorded future. We replace only the object path with a translated alternative, then use a detector that receives neither path to measure whether the generated object moves in response. Figure 2 summarizes this controlled design with examples from real videos.

3.1 Benchmark construction

The evaluation population contains 200 broadcast windows, with 50 per sport, selected before detection or generation by a fixed hash order and round-robin source traversal. We exclude sources in the adaptation-training manifests, colliding clip identifiers, and temporal overlap within a source.

The two halves of a SoccerNet-style match share one fixture identity, and every selected window is decoded again from its source. The soccer windows come from one match, while basketball, badminton, and table tennis span 50, 18, and 13 recordings. Soccer uncertainty is therefore conditional on that match. The source audit establishes separation from documented adaptation data, although unknown backbone pretraining and re-edits of the same event cannot be ruled out. The continuation-training population contains 400 additional windows, with 100 per sport and no evaluation source identity.

Each window contains 26 frames at 512×896 pixels and 25 frames per second, with five context frames followed by a 21-frame future. Replacement machine commands are available in 40 soccer, 37 basketball, 42 badminton, and 48 table-tennis windows. Items without commands, close-ups, transitions, occlusions, and uncertain localizations remain in the population, preventing detector success or generated outcomes from redefining it.

The artifacts retain source identities, selection decisions, manifest hashes, commands, detector provenance, measurements, and license records. Having fixed the population and temporal task, we next define the path intervention.

3.2 Paired path intervention

Let $c^0_{it} \in [0,1]^2$ denote the object command for item i at future frame t. We construct c^1_{it} by shifting its horizontal coordinate by $134/895$, approximately 15% of the image width. A fixed item hash selects the direction independently of scene content and results. Both paths retain the same temporal support after positions outside the image are removed.

The selected path defines an object mask and Gaussian trajectory heat map. Only these channels change between arms. The context, prompt, playing-surface and player controls, optical flow, model, sampler, guidance, and random noise remain identical. The pair therefore isolates the effect of replacing the instructed path.

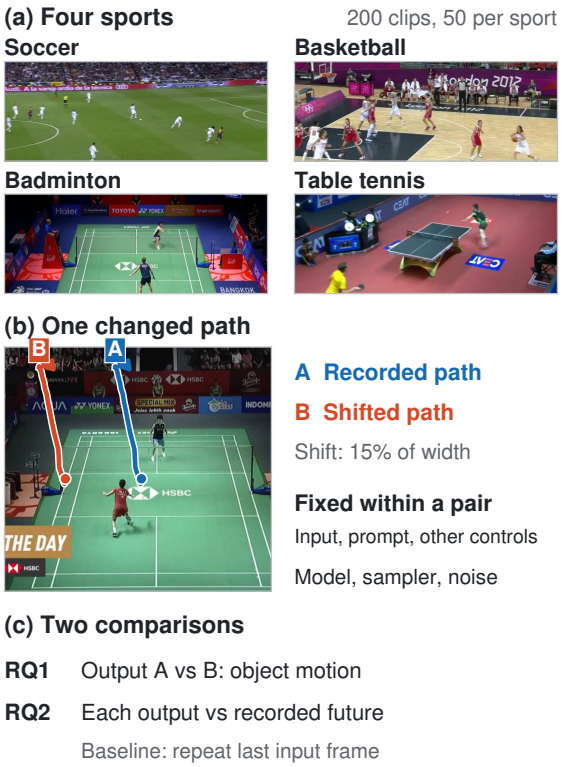

Figure 2: **One intervention, two distinct tests.** (a) Each broadcast window has five input and 21 future frames. (b) Machine-annotated path A is shifted horizontally by 134 pixels to form B. Only the object controls change within a pair. (c) RQ1 compares generated object motion using a detector that sees neither command. RQ2 compares each output with the recorded future, using frame repetition as a baseline. Future similarity alone does not establish path following.

The translated path may conflict with player motion, camera geometry, or sport rules. Following it establishes sensitivity to the supplied control rather than physical correctness, while an unchanged realistic continuation does not establish control.

A sport-specific detector returns positions y^0_{it} and y^1_{it} without access to either command. Coordinates are normalized by image width minus one and height minus one. At the primary confidence threshold of 0.5, we neither interpolate missing positions nor use commands to select detections. The detector, preprocessing, and three-frame inference protocol remain fixed across arms.

3.3 Response, proximity, and observation coverage

For a frame with commands and detections in both arms, we compare each output with both possible paths:

$$P_{it} = \frac{\|y^0_{it} - c^1_{it}\|_2 - \|y^0_{it} - c^0_{it}\|_2 + \|y^1_{it} - c^0_{it}\|_2 - \|y^1_{it} - c^1_{it}\|_2}{2\|c^1_{it} - c^0_{it}\|_2}. \tag{1}$$

By the triangle inequality, $P_{it} \in [-1, 1]$. Exact following gives one, identical detected outputs give zero regardless of their location, and negative values favor the alternative assignment. The cancellation prevents an ordinary continuation from receiving positive credit merely for lying near both commands.

Preference measures assigned-path separation rather than complete localization accuracy. We accompany it with assigned-command distance, success within normalized distances 0.02 and 0.05, and horizontal response gain $(y^1_{it,x} - y^0_{it,x})/(c^1_{it,x} - c^0_{it,x})$. Missing detections receive no success credit. A gain of one denotes the commanded horizontal displacement and zero denotes no measured response. Preference remains primary because gain does not require proximity to the assigned paths.

Let N_i be the number of commanded frames, O_i the subset observed in both outputs, and $n_i = |O_i|$. The conditional item score averages P_{it} over O_i, while joint coverage is n_i/N_i. To expose the effect of missing observations, we also report the worst-case identification interval

$$\left[\frac{\sum_{t\in O_i} P_{it} - (N_i - n_i)}{N_i}, \frac{\sum_{t\in O_i} P_{it} + (N_i - n_i)}{N_i}\right]. \tag{2}$$

The interval assigns every missing frame each value allowed by the score range. It is not a confidence interval or a correction for false localization. An item without a joint detection has no point estimate and retains $[-1, 1]$.

3.4 Real-video calibration and uncertainty

Because the commands and observations use the same detector family, comparing commands with unchanged source frames would be circular. We instead translate the real future by the known command displacement and ask whether the pipeline recovers that change. An unchanged pair provides a zero-response control. These controls test transformation response and coverage rather than semantic accuracy, which a separate blinded point inspection probes in Appendix C.

We report sports separately. Frame scores and then generated seeds are averaged within items. The factorial comparison uses the common seed-zero population across its four arms. Descriptive 95% intervals use 2,000 bootstrap draws of source recordings with paired conditions retained, except soccer, whose single-match analysis resamples clips. Threshold analyses at 0.7 and 0.9 report the loss of observations. Pixel response and similarity to the recorded future remain separate because neither establishes path response.

4 Experimental Setup

In this section, we first describe the experimental designs for our two research questions in Section 4.1, and then our training and evaluation setup in Section 4.2.

4.1 Experimental designs

We define two key experimental design choices. The first corresponds to response to intervention (RQ1), and the second to future similarity (RQ2). We vary instructed paths, training annotations, and control alignment independently under RQ1, and comparison references under RQ2. The overview is in Figure 2, with details in Appendix A.

Matched vs. shifted paths. To investigate RQ1, we construct two path regimes: matched and shifted. The matched path is estimated from the recorded future, while the shifted path translates it horizontally by 134 pixels, approximately 15% of the image width, in a direction set by a fixed item hash. Both paths have identical temporal support. Within each pair, the context, prompt, other controls, model, sampling procedure, and noise remain fixed, isolating response to the changed path.

We apply the intervention under two training-annotation regimes: original heuristic object paths developed for soccer imagery and machine paths from sport-specific WASB detectors (Tarashima et al., 2023). Identical evaluation paths isolate the effect of training annotations. We also construct two control-alignment regimes. The flattened regime resizes a flattened control sequence to the backbone token count, which can mix space and time. The explicit-grid regime preserves the causal temporal groups and spatial grid before flattening. Crossing both factors yields four conditions from the same adapted 2B checkpoint, allowing their main effects and interaction to be measured.

We train every condition on the same 400 windows, with 100 per sport and no evaluation source identity. Fixed hash ordering, round-robin selection, clip order, training seed, and update budget keep the

training data and procedure comparable. We then transfer the sport-specific, explicit-grid condition to an existing 14B recipe and vary resolution, generation seed, and object-channel removal. Since the recipes differ in training history and adapter configuration, this transfer tests robustness rather than model capacity.

Recorded vs. repeated-frame references. Under RQ2, we evaluate future similarity with two references: the recorded continuation and a static baseline that repeats the final context frame. We use the same 200 evaluation windows, with 50 per sport, and reuse the matched-path and shifted-path generations from RQ1 without selecting outputs. Framewise LPIPS (Zhang et al., 2018) measures whether each generation is closer to the recorded future than the static baseline and whether this advantage changes with the path. Because the recording is one plausible future, this design measures reference similarity rather than adherence.

4.2 Training and evaluation setup

Training. Given five context frames $x_{0:4}$, a sport prompt p, and future controls C, the model generates 21 future frames $\hat{x}_{5:25}$ at 25 frames per second. Ten spatial channels encode surface lines, player and role masks, an object mask, a trajectory heat map, and optical flow. The object is a ball except in badminton, where it is a shuttlecock. Availability and sport identity form additional input planes, and controls are stored at 128×224 pixels.

A three-dimensional convolutional encoder maps these inputs to features, which zero-initialized residual projections inject into transformer blocks. We continue the existing adapters and LoRA weights (Hu et al., 2022).

Each arm receives 1,000 updates on the same 400 clips using AdamW (Loshchilov & Hutter, 2019) with learning rate 10^{-4}, batch size one, and gradient clipping at norm one. The backbone and text encoder remain frozen while LoRA and adapter weights are updated. Training uses five context and 20 target frames at 256×448 pixels. For $z_t = (1-t)z + t\epsilon$, with $t \sim \mathcal{U}(0,1)$ and $\epsilon \sim \mathcal{N}(0, I)$, the objective is mean squared rectified-flow velocity error (Liu et al., 2023) on the future latent elements marked by M, normalized by $\|M\|_1$. Clean context latents use a conditioning mask, and adapter gradients are verified before the first update. We report the fixed final checkpoint without adherence-based selection.

Evaluation. Each sports condition uses the 200 fixed windows, paired commands, and a 20-step UniPC multistep sampler (Zhao et al., 2023) with guidance scale 1.0. At 256×448, we evaluate both original recipes, all four 2B continuation arms, and the corrected 14B continuation. The corrected 2B and 14B recipes also receive a second generated seed and an object-channel-removal condition. At 512×896, both original and corrected recipes receive paired generation at seed zero. The sports grid contains 5,600 generated futures and appears in Appendix A.

All outputs share a 512×896 evaluation canvas, with reduced-resolution videos expanded by area resampling. The factorial comparison uses seed zero across all arms. No critic or score selects samples. Following Section 3, we report sports separately, retain missing observations, and use descriptive pointwise 95% bootstrap intervals with paired conditions kept together. The comparisons are descriptive rather than independent significance tests.

5 Response to Intervention (RQ1)

We first study whether changing only the instructed path of a ball or shuttlecock changes its generated motion. For the same input video, we generate two continuations under otherwise identical conditions. We measure this change with an automatic object detector, which we first verify on real videos in which the object is moved by a known amount. We then vary the training paths and how those paths enter the model, before extending the analysis to further adaptation recipes, resolutions, and control inputs in Appendix A.5 and to non-sports domains in Section 7.

The response measure detects a known change in object position. Before applying the measure to generated videos, we test it on real videos for which the expected result is known. We compare each video with an identical copy and with a copy in which every frame is shifted horizontally by the instructed distance. Identical copies produce a response score of zero, while shifted copies produce scores between 0.956 and 0.988 across the four sports (Appendix Table 2). The measure therefore

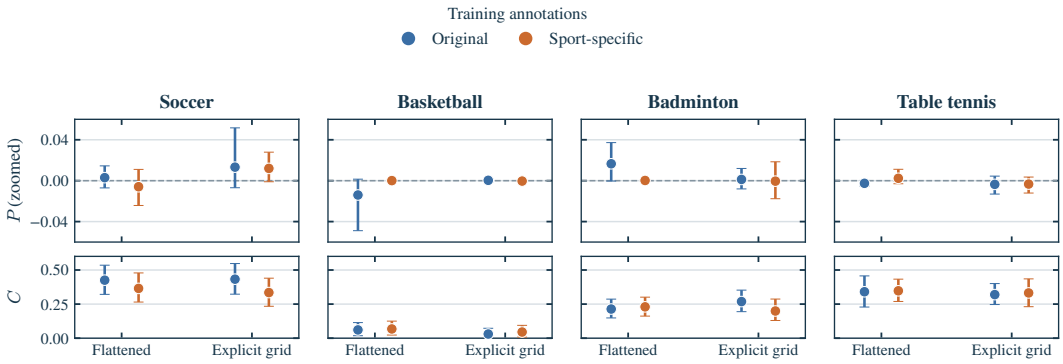

Figure 3: **Response across training annotations and control alignments.** The top row shows preference P on a shared scale zoomed around zero, while the bottom row shows joint detection coverage C. The theoretical range of P is $[-1, 1]$: zero denotes no paired response and one denotes exact following. Coverage is the fraction of commanded frames detected in both outputs. Points show seed-zero estimates and whiskers show 95% intervals. Open markers denote estimates whose intervals are unavailable. All four arms share the initial checkpoint, training data, and update budget.

Table 1: **Effects of training-path source and control alignment.** At seed zero, ΔS averages matched- and shifted-path success within normalized distance 0.05. Missing detections count as failures. Intervals are 95%, and n counts clips with commanded frames.

Sport	Annotation ΔS	Alignment ΔS	Interaction ΔS	n
Soccer	-0.006 [-0.027, 0.013]	-0.010 [-0.031, 0.008]	-0.022 [-0.069, 0.026]	40
Basketball	0.002 [-0.006, 0.012]	-0.008 [-0.026, 0.004]	-0.002 [-0.038, 0.022]	37
Badminton	-0.007 [-0.021, 0.005]	-0.006 [-0.019, 0.005]	-0.018 [-0.043, 0.004]	42
Table tennis	0.002 [-0.007, 0.011]	0.001 [-0.009, 0.011]	-0.006 [-0.024, 0.010]	48

distinguishes no change from a known change in object position. This calibration tests the response calculation, while a separate blinded audit evaluates object-localization accuracy in Appendix C.

Changing the training paths or how they are given to the model does not improve path following. We cross two sources of training paths with two ways of passing those paths to the model (Figure 3). For the same input, changing only the path assigned to the ball or shuttlecock does not consistently move the generated object in any variant. The conclusion is unchanged when missing detections count as failures and success requires placing the object within normalized distance 0.05 of its instructed position (Table 1). The strict four-variant response is unavailable for basketball because no frame is detected in every required video, while the remaining sports show no reliable improvement from either training change.

Takeaway 1: The evaluation recovers known translations, but the tested training variants show no reliable positive response to a changed object path. Neither sport-specific training paths nor explicit control-grid alignment consistently improves path following.

6 Future Similarity (RQ2)

We investigate RQ2: whether generated videos resemble the recorded future and whether this resemblance changes with the instructed path. Existing evaluations assess visual and temporal quality (Liu et al., 2024; Huang et al., 2024) or similarity to a recorded continuation (Chen et al., 2023). Since the input already suggests a future, similarity alone cannot establish control. Using LPIPS (Zhang et al., 2018), we compare native-resolution generation with a repeated-frame baseline and compare matched with shifted paths. Appendix A reports the full grid.

Native-resolution generations usually resemble the recorded future more than a repeated frame. All sixteen point estimates in Figure 4 (top) are positive, and thirteen intervals exclude zero.

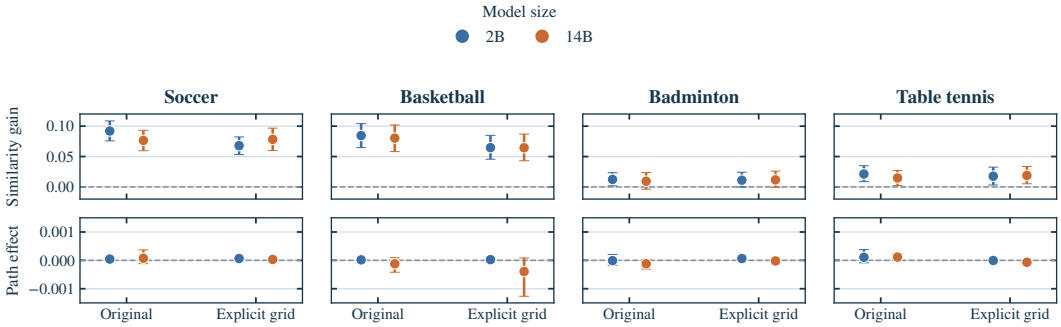

Figure 4: **Similarity to the recorded future at native resolution.** The top row shows repeated-frame LPIPS minus generated LPIPS, where positive values favor generation. The bottom row shows shifted-path LPIPS minus matched-path LPIPS, where zero means that changing the path leaves future similarity unchanged. Points and whiskers show means and pointwise 95% intervals over 50 clips per sport, with source-level resampling except for single-match soccer (Section 3).

LPIPS improves by 0.065 to 0.092 for soccer and basketball and by 0.009 to 0.021 for badminton and table tennis. At reduced resolution, however, the corrected configurations are less similar than the repeated-frame baseline in every sport.

Changing the instructed path barely changes future similarity. For paired videos that differ only in the assigned path, fourteen of sixteen intervals in Figure 4 (bottom) include zero, and the two positive differences are below 0.00012 LPIPS. Although mean pixel differences range from 0.005 to 0.012, matched- and shifted-path videos remain almost equally similar to the recorded future. Figure 1(c) summarizes this distinction for the corrected 2B model across 200 clips: generation reduces mean LPIPS from 0.223 for repetition to 0.183 for either path condition, an improvement of about 18%. Thus the model predicts the recorded future better, but supplying its recorded path provides almost no additional benefit over an altered path. Because LPIPS compares the full video, this observation alone does not establish whether the ball follows a command, which is tested separately under RQ1.

Takeaway 2: Native-resolution generations usually resemble the recorded future more than a repeated frame, yet this similarity barely changes with the instructed path. Future similarity and motion control are distinct capabilities.

7 Generalization Beyond Sports

Our sports experiments reveal two findings: changing the instructed path does not reliably change object motion (RQ1), while generated continuations can resemble the recorded future more closely than a static baseline (RQ2). We next examine whether these findings extend to animals and road vehicles, whose appearances, movements, and surroundings differ from balls and shuttlecocks. We use real videos from DAVIS (Pont-Tuset et al., 2017), comprising 13 animal windows from seven source videos and 10 vehicle windows from six sources. Published instance masks define the object paths independently of the detector used for evaluation. Both sports-adapted checkpoints remain frozen, with sports-specific controls disabled and no additional training. We retain the paired-path intervention, native resolution, and two sampling seeds, producing 184 videos. Appendix I gives the selection, setup, and complete results.

Object-motion control beyond sports. To investigate RQ1, we change only the instructed animal or vehicle path while keeping the context and sampling noise fixed. Neither recipe shows a reliably positive response in either domain (Figure 5, top). All four preference intervals include zero, even though joint detection coverage exceeds 97%. The same evaluation registers known translations in real videos, with mean preference above 0.8 in both domains (Appendix Table 96). Thus, limited measured response also occurs outside sports, where missing detections are less frequent. This does not establish correct instance tracking in every frame.

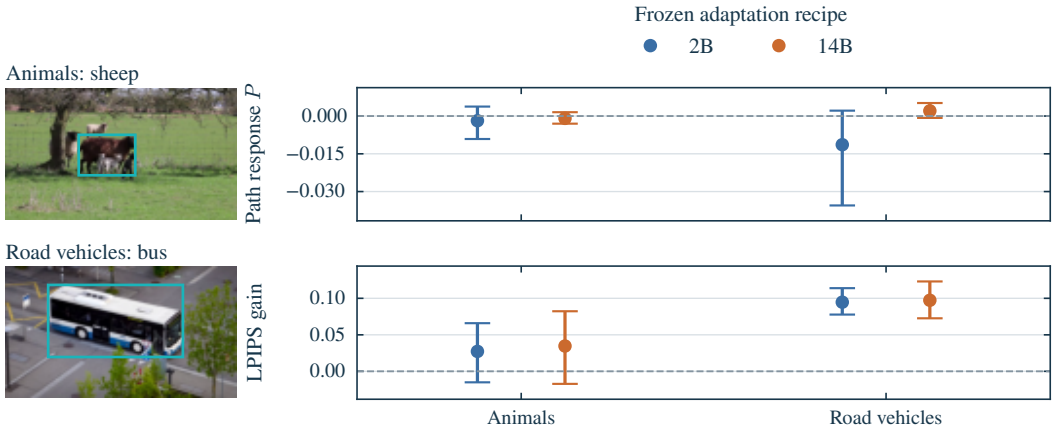

Figure 5: **Path response and future similarity beyond sports.** Left: real DAVIS context frames with annotated target boxes. Right: path preference P (top, zoomed from its full range $[-1, 1]$) and LPIPS improvement over repeating the last context frame (bottom). Zero in the top row denotes no paired path response, while positive values in the bottom row favor generation. Neither recipe shows reliably positive path response, although both improve future similarity for vehicles. Whiskers show pointwise 95% intervals from resampling source videos after averaging seeds within windows.

Future similarity beyond sports. For RQ2, we compare the same generations with the recorded future and repeated-frame baseline. Both recipes improve reference similarity for road vehicles, with positive gain intervals (Figure 5, bottom). Animal gains are smaller and their intervals include zero. Changing the instructed path has little effect on LPIPS in either domain (Appendix Table 98).

Takeaway 3: Frozen sports-adapted models show limited path response on animals and road vehicles. Improved future similarity extends to vehicles, but remains uncertain for animals.

8 Limitations and Responsible Use

Our claims concern measured response under a detector that can miss or mislocalize small objects, which calibration, missingness bounds, and a blinded audit only partially address. The study covers four sports and a smaller transfer test on animals and road vehicles, with short futures, two model configurations, one continuation-training seed, and a fixed budget, so it does not establish an intrinsic limitation of diffusion models. The translated paths need not describe feasible play, the recording is only one possible future, and no human preference study was conducted. Footage permissions, identification risks, and responsible-use requirements appear in Appendix H.

9 Conclusion

SportWorldBench separates two questions that plausible video continuations can obscure: whether a model responds to an instructed path and whether it predicts the recorded future. Known translations yield near-unit response in our evaluation, yet generated objects show no reliable positive response when their instructed paths change. This finding persists across the tested annotations, control alignments, training recipes, resolutions, and sampling seeds. At native resolution, generations often resemble the recorded future more than a repeated input frame, with nearly identical similarity under recorded and altered paths. The vehicle transfer test exhibits the same separation.

These results motivate evaluating motion control through paired interventions, alongside prediction quality, detection coverage, and missing-observation bounds. For future methods, the central objective is to make a changed spatial command produce the corresponding change in generated object motion. Improved visual prediction alone does not demonstrate that capability.

REPRODUCIBILITY STATEMENT

The accompanying experiment artifacts contain the fixed selection, source and model hashes, command tensors, training configurations, per-item measurements, and scripts that generate the tables and figures. Appendix F specifies the execution and validation procedures. Machine annotations and detector failures are retained in the reports. Access to source footage remains subject to its original permissions.

ETHICS STATEMENT

Section 8 describes the scope of the claims, machine-annotation limitations, and broadcast-content considerations. We do not claim that the image-space interventions describe physically feasible play.

AI USE STATEMENT

An AI coding assistant supported experiment design, implementation, debugging, literature checking, analysis, and manuscript drafting. It also performed the explicitly labeled model-assisted point inspection in Appendix C. That inspection is not a human study or independent ground truth. The experiment outputs, provenance records, and numerical reports are retained for verification.

REFERENCES

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. URL https://arxiv.org/abs/2311.15127.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. SEINE: Short-to-long video diffusion model for generative transition and prediction. *arXiv preprint arXiv:2310.20700*, 2023. URL https://arxiv.org/abs/2310.20700.

Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. SoccerNet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018. URL https://arxiv.org/abs/1804.04527.

Thomas Gossard, Filip Radovic, Andreas Ziegler, and Andreas Zell. BlurBall: Joint ball and motion blur estimation for table tennis ball tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2026. URL https://arxiv.org/abs/2509.18387.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2204.03458.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.17982.

Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. PEEKABOO: Interactive video generation via masked-diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8079–8088, 2024. URL https://arxiv.org/abs/2312.07509.

Yujin Jeong, Arnas Uselis, Iro Laina, Seong Joon Oh, and Anna Rohrbach. When do diffusion models learn to generate multiple objects? In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.00273.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.05954.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025a. URL https://arxiv.org/abs/2502.20694.

Quanhao Li, Zhen Xing, Rui Wang, Hui Zhang, Qi Dai, and Zuxuan Wu. MagicMotion: Controllable video generation with dense-to-sparse trajectory guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025b. URL https://arxiv.org/abs/2503.16421.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.03003.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2310.11440.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

NVIDIA. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025a. URL https://arxiv.org/abs/2501.03575.

NVIDIA. World simulation with video foundation models for physical AI. *arXiv preprint arXiv:2511.00062*, 2025b. URL https://arxiv.org/abs/2511.00062.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. URL https://arxiv.org/abs/1704.00675.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. URL https://arxiv.org/abs/1506.01497.

Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Motion-I2V: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH Conference Papers*, 2024. URL https://arxiv.org/abs/2401.15977.

Shuhei Tarashima, Muhammad Abdul Haq, Yushan Wang, and Norio Tagawa. Widely applicable strong baseline for sports ball detection and tracking. In *British Machine Vision Conference*, 2023. URL https://arxiv.org/abs/2311.05237.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges. *arXiv preprint arXiv:1812.01717*, 2018. URL https://arxiv.org/abs/1812.01717.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH Conference Papers*, 2024. URL https://arxiv.org/abs/2312.03641.

Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. DragAnything: Motion control for anything using entity representation. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2403.07420.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.06072.

Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Ming Gong, and Nan Duan. DragNUWA: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. URL https://arxiv.org/abs/2308.08089.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2302.05543.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. URL https://arxiv.org/abs/1801.03924.

Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2407.21705.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2302.04867.

A Complete Experimental Grid and Results

The tables below retain the entire fixed campaign, including every sport, the original checkpoints, all four matched continuation arms, both rendering resolutions, and the generated-seed and object-removal checks. The internal label `clean` in artifact names denotes sport-specific machine annotations, not verified ground truth. The internal label `causal` denotes explicit control-grid alignment. All primary checkpoints are the fixed final 1,000-update continuations. Intermediate checkpoints were saved for recovery and engineering diagnosis and do not select the reported result.

A.1 Real-video calibration

Table 2: **Known-translation calibration.** Preference P and joint detection coverage C for translated and unchanged real-video controls at native and resampled resolution. Intervals are 95%. This tests transformation response rather than independent semantic accuracy.

Sport	Scale	Translated P (95% CI)	C	Unchanged P (95% CI)	C
Soccer	Native	0.978 [0.960, 0.990]	0.90	0.000 [0.000, 0.000]	1.00
	Resampled	0.919 [0.884, 0.951]	0.67	0.000 [0.000, 0.000]	0.70
Basketball	Native	0.988 [0.965, 0.998]	0.90	0.000 [0.000, 0.000]	1.00
	Resampled	0.983 [0.955, 0.996]	0.76	0.000 [0.000, 0.000]	0.81
Badminton	Native	0.978 [0.946, 0.997]	0.97	0.000 [0.000, 0.000]	1.00
	Resampled	0.969 [0.932, 0.992]	0.93	0.000 [0.000, 0.000]	0.97
Table tennis	Native	0.956 [0.933, 0.974]	0.91	0.000 [0.000, 0.000]	1.00
	Resampled	0.924 [0.892, 0.952]	0.85	0.000 [0.000, 0.000]	0.92

A.2 CONFIGURATION GRID

Table 3: Complete fixed generation grid. Every row contains all 50 selected items in each of the four sports. All rows use 20 sampling steps and guidance 1.0. The total is 5,600 videos.

Recipe	Render size	Object condition	Seeds	Videos
Original 2B	256 by 448	matched	1	200
Original 2B	256 by 448	shifted	1	200
Original 2B	512 by 896	matched	1	200
Original 2B	512 by 896	shifted	1	200
Original 14B	256 by 448	matched	1	200
Original 14B	256 by 448	shifted	1	200
Original 14B	512 by 896	matched	1	200
Original 14B	512 by 896	shifted	1	200
2B grid / sport labels	256 by 448	matched	2	400
2B grid / sport labels	256 by 448	shifted	2	400
2B grid / sport labels	256 by 448	removed	1	200
2B grid / sport labels	512 by 896	matched	1	200
2B grid / sport labels	512 by 896	shifted	1	200
2B grid / original labels	256 by 448	matched	1	200
2B grid / original labels	256 by 448	shifted	1	200
2B flat / sport labels	256 by 448	matched	1	200
2B flat / sport labels	256 by 448	shifted	1	200
2B flat / original labels	256 by 448	matched	1	200
2B flat / original labels	256 by 448	shifted	1	200
14B grid / sport labels	256 by 448	matched	2	400
14B grid / sport labels	256 by 448	shifted	2	400
14B grid / sport labels	256 by 448	removed	1	200
14B grid / sport labels	512 by 896	matched	1	200
14B grid / sport labels	512 by 896	shifted	1	200

A.3 COMMAND PREFERENCE, LOCALIZATION, AND MISSINGNESS

Table 4: **Path response at** 256×448**.** Results average two seeds per clip. P is preference, C is joint coverage, and S_0, S_1 are success at distance 0.05. Bounds retain missing detections. Intervals are 95%, and n counts measured clips.

Recipe	Sport	P (95% CI)	C	S_0	S_1	Missingness bounds	n
2B	Soccer	0.002 [-0.005, 0.009]	0.34	0.12	0.02	[-0.66, 0.67]	30
	Basketball	0.000 [-0.001, 0.000]	0.05	0.01	0.00	[-0.95, 0.95]	9
	Badminton	-0.002 [-0.012, 0.007]	0.18	0.08	0.00	[-0.82, 0.82]	32
	Table tennis	-0.003 [-0.009, 0.002]	0.32	0.08	0.01	[-0.68, 0.68]	42
14B	Soccer	-0.017 [-0.037, -0.001]	0.43	0.16	0.01	[-0.57, 0.57]	34
	Basketball	0.001 [-0.001, 0.004]	0.02	0.00	0.00	[-0.98, 0.98]	5
	Badminton	-0.006 [-0.016, 0.000]	0.22	0.08	0.00	[-0.78, 0.78]	32
	Table tennis	-0.005 [-0.016, 0.006]	0.35	0.07	0.01	[-0.65, 0.65]	45

Table 5: Original 2B, render height 256. Preference averages generated seeds within clips when two seeds are present. C is joint coverage, and S_0, S_1 are command-proximity success at 0.05 in matched and shifted outputs. Bounds retain missing commanded frames. The last column counts clips with conditional preference.

Sport	P [95% CI]	C	S_0	S_1	Missingness bounds	n
Soccer	-0.002 [-0.016, 0.008]	0.390	0.165	0.010	[-0.611, 0.609]	28
Basketball	0.001 [0.000, 0.002]	0.079	0.010	0.000	[-0.921, 0.921]	7
Badminton	0.001 [-0.008, 0.009]	0.320	0.111	0.001	[-0.681, 0.679]	33
Table tennis	-0.017 [-0.037, -0.002]	0.359	0.071	0.007	[-0.647, 0.635]	39

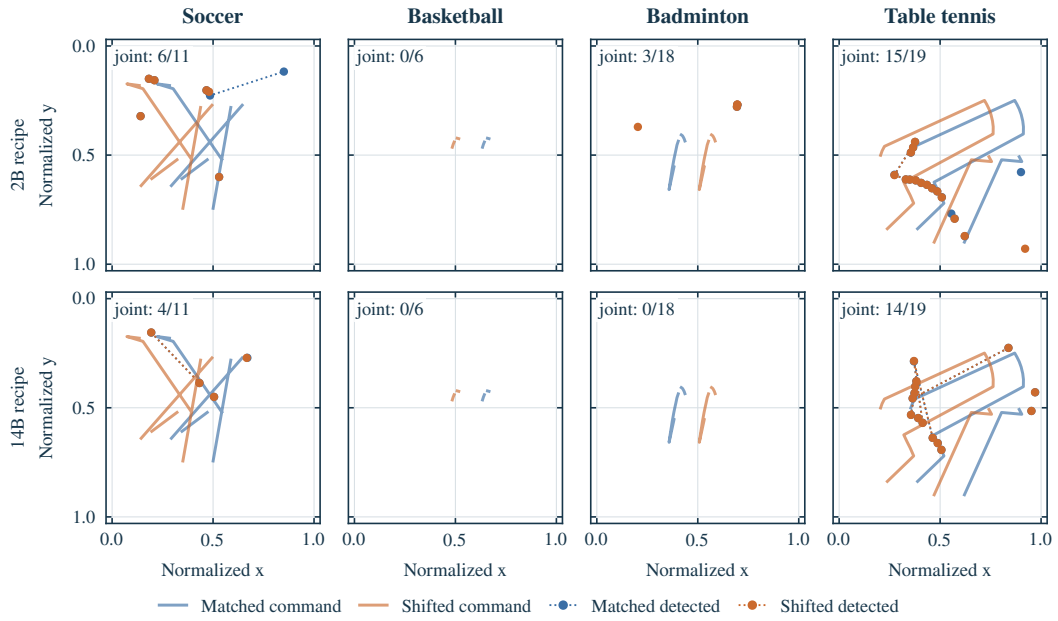

Figure 6: **Fixed illustrative trajectories across all four sports.** For each sport, a fixed hash selects one clip among those with at least five commanded frames, independently of generated outcomes. The two corrected-grid recipes use the same selected clips, reduced-resolution generation, and seed zero. Solid paths show commands and dotted paths with markers show detected positions on commanded frames. Paths connect consecutive observations and break at missing frames. The vertical axis follows image coordinates. Each panel gives the number of frames observed in both generated arms out of all commanded frames. These are machine trajectories and illustrative cases, not ground truth or a representative sample.

Table 6: Original 2B, render height 512. Preference averages generated seeds within clips when two seeds are present. C is joint coverage, and S_0, S_1 are command-proximity success at 0.05 in matched and shifted outputs. Bounds retain missing commanded frames. The last column counts clips with conditional preference.

Sport	P [95% CI]	C	S_0	S_1	Missingness bounds	n
Soccer	0.015 [-0.004, 0.047]	0.355	0.211	0.014	[-0.641, 0.649]	24
Basketball	0.000 [-0.001, 0.000]	0.098	0.054	0.000	[-0.902, 0.902]	7
Badminton	-0.004 [-0.024, 0.015]	0.589	0.271	0.000	[-0.417, 0.405]	38
Table tennis	0.005 [-0.006, 0.023]	0.523	0.181	0.006	[-0.477, 0.477]	46

Table 7: Original 14B, render height 256. Preference averages generated seeds within clips when two seeds are present. C is joint coverage, and S_0, S_1 are command-proximity success at 0.05 in matched and shifted outputs. Bounds retain missing commanded frames. The last column counts clips with conditional preference.

Sport	P [95% CI]	C	S_0	S_1	Missingness bounds	n
Soccer	-0.023 [-0.047, -0.003]	0.347	0.056	0.006	[-0.660, 0.646]	32
Basketball	0.000 [-0.033, 0.032]	0.080	0.011	0.003	[-0.923, 0.918]	13
Badminton	-0.008 [-0.036, 0.009]	0.100	0.008	0.000	[-0.901, 0.899]	20
Table tennis	-0.023 [-0.058, 0.004]	0.221	0.023	0.009	[-0.782, 0.776]	45

Table 8: Original 14B, render height 512. Preference averages generated seeds within clips when two seeds are present. C is joint coverage, and S_0, S_1 are command-proximity success at 0.05 in matched and shifted outputs. Bounds retain missing commanded frames. The last column counts clips with conditional preference.

Sport	P [95% CI]	C	S_0	S_1	Missingness bounds	n
Soccer	0.000 [-0.009, 0.007]	0.577	0.256	0.006	[-0.424, 0.422]	36
Basketball	0.000 [-0.001, 0.000]	0.076	0.032	0.000	[-0.924, 0.924]	10
Badminton	0.001 [-0.002, 0.005]	0.417	0.207	0.007	[-0.582, 0.584]	33
Table tennis	0.000 [-0.023, 0.030]	0.524	0.146	0.008	[-0.478, 0.475]	43

Table 9: 2B grid / sport labels, render height 256. Preference averages generated seeds within clips when two seeds are present. C is joint coverage, and S_0, S_1 are command-proximity success at 0.05 in matched and shifted outputs. Bounds retain missing commanded frames. The last column counts clips with conditional preference.

Sport	P [95% CI]	C	S_0	S_1	Missingness bounds	n
Soccer	0.002 [-0.005, 0.009]	0.336	0.125	0.015	[-0.662, 0.666]	30
Basketball	0.000 [-0.001, 0.000]	0.045	0.009	0.004	[-0.955, 0.955]	9
Badminton	-0.002 [-0.012, 0.007]	0.184	0.075	0.001	[-0.817, 0.815]	32
Table tennis	-0.003 [-0.009, 0.002]	0.320	0.077	0.009	[-0.681, 0.679]	42

Table 10: 2B grid / sport labels, render height 512. Preference averages generated seeds within clips when two seeds are present. C is joint coverage, and S_0, S_1 are command-proximity success at 0.05 in matched and shifted outputs. Bounds retain missing commanded frames. The last column counts clips with conditional preference.

Sport	P [95% CI]	C	S_0	S_1	Missingness bounds	n
Soccer	0.001 [-0.004, 0.007]	0.394	0.184	0.007	[-0.606, 0.606]	29
Basketball	0.000 [0.000, 0.001]	0.042	0.030	0.000	[-0.958, 0.958]	7
Badminton	0.003 [-0.002, 0.010]	0.338	0.174	0.001	[-0.660, 0.665]	33
Table tennis	-0.006 [-0.014, 0.000]	0.446	0.119	0.005	[-0.558, 0.551]	43

Table 11: 2B grid / original labels, render height 256. Preference averages generated seeds within clips when two seeds are present. C is joint coverage, and S_0, S_1 are command-proximity success at 0.05 in matched and shifted outputs. Bounds retain missing commanded frames. The last column counts clips with conditional preference.

Sport	P [95% CI]	C	S_0	S_1	Missingness bounds	n
Soccer	0.013 [-0.007, 0.052]	0.433	0.175	0.008	[-0.561, 0.574]	30
Basketball	0.000 [-0.001, 0.002]	0.031	0.020	0.000	[-0.969, 0.969]	5
Badminton	0.001 [-0.008, 0.012]	0.269	0.092	0.002	[-0.732, 0.731]	30
Table tennis	-0.004 [-0.013, 0.004]	0.320	0.067	0.007	[-0.681, 0.679]	40

Table 12: 2B flat / sport labels, render height 256. Preference averages generated seeds within clips when two seeds are present. C is joint coverage, and S_0, S_1 are command-proximity success at 0.05 in matched and shifted outputs. Bounds retain missing commanded frames. The last column counts clips with conditional preference.

Sport	P [95% CI]	C	S_0	S_1	Missingness bounds	n
Soccer	-0.006 [-0.024, 0.011]	0.365	0.176	0.016	[-0.638, 0.632]	26
Basketball	0.000 [0.000, 0.000]	0.068	0.040	0.002	[-0.932, 0.932]	8
Badminton	0.000 [0.000, 0.001]	0.230	0.090	0.001	[-0.770, 0.770]	30
Table tennis	0.002 [-0.003, 0.011]	0.348	0.075	0.002	[-0.651, 0.654]	41

Table 13: 2B flat / original labels, render height 256. Preference averages generated seeds within clips when two seeds are present. C is joint coverage, and S_0, S_1 are command-proximity success at 0.05 in matched and shifted outputs. Bounds retain missing commanded frames. The last column counts clips with conditional preference.

Sport	P [95% CI]	C	S_0	S_1	Missingness bounds	n
Soccer	0.003 [-0.007, 0.014]	0.426	0.167	0.015	[-0.573, 0.576]	30
Basketball	-0.014 [-0.049, 0.001]	0.061	0.035	0.000	[-0.941, 0.938]	8
Badminton	0.017 [0.000, 0.037]	0.214	0.085	0.003	[-0.783, 0.790]	28
Table tennis	-0.003 [-0.005, 0.000]	0.341	0.058	0.008	[-0.660, 0.658]	41

Table 14: 14B grid / sport labels, render height 256. Preference averages generated seeds within clips when two seeds are present. C is joint coverage, and S_0, S_1 are command-proximity success at 0.05 in matched and shifted outputs. Bounds retain missing commanded frames. The last column counts clips with conditional preference.

Sport	P [95% CI]	C	S_0	S_1	Missingness bounds	n
Soccer	-0.017 [-0.037, -0.001]	0.431	0.164	0.011	[-0.573, 0.565]	34
Basketball	0.001 [-0.001, 0.004]	0.015	0.002	0.000	[-0.985, 0.985]	5
Badminton	-0.006 [-0.016, 0.000]	0.221	0.078	0.000	[-0.781, 0.776]	32
Table tennis	-0.005 [-0.016, 0.006]	0.346	0.069	0.007	[-0.654, 0.653]	45

Table 15: 14B grid / sport labels, render height 512. Preference averages generated seeds within clips when two seeds are present. C is joint coverage, and S_0, S_1 are command-proximity success at 0.05 in matched and shifted outputs. Bounds retain missing commanded frames. The last column counts clips with conditional preference.

Sport	P [95% CI]	C	S_0	S_1	Missingness bounds	n
Soccer	-0.021 [-0.068, 0.000]	0.375	0.204	0.012	[-0.629, 0.621]	31
Basketball	0.000 [0.000, 0.001]	0.049	0.017	0.000	[-0.951, 0.951]	7
Badminton	0.001 [0.000, 0.002]	0.288	0.157	0.008	[-0.712, 0.713]	30
Table tennis	-0.008 [-0.042, 0.018]	0.377	0.133	0.010	[-0.624, 0.622]	40

A.4 Matched factorial and sensitivity contrasts

Table 16: Factorial annotation contrast at seed zero. Preference uses frames detected in all four arms. Command success uses the full fixed command support. Missingness bounds combine the full-support arm bounds, and are not confidence intervals.

Sport	Preference [95% CI]	Common coverage	Matched success Δ	Missingness bounds
Soccer	-0.007 [-0.042, 0.021]	0.191	-0.018 [-0.060, 0.020]	[-1.224, 1.218]
Basketball	n/a	0.000	0.001 [-0.015, 0.018]	[-1.897, 1.899]
Badminton	-0.036 [-0.081, -0.006]	0.108	-0.013 [-0.041, 0.011]	[-1.546, 1.542]
Table tennis	-0.010 [-0.049, 0.009]	0.146	0.007 [-0.012, 0.023]	[-1.329, 1.331]

Table 17: Factorial alignment contrast at seed zero. Preference uses frames detected in all four arms. Command success uses the full fixed command support. Missingness bounds combine the full-support arm bounds, and are not confidence intervals.

Sport	Preference [95% CI]	Common coverage	Matched success Δ	Missingness bounds
Soccer	0.023 [-0.007, 0.057]	0.191	-0.019 [-0.059, 0.015]	[-1.214, 1.228]
Basketball	n/a	0.000	-0.018 [-0.052, 0.004]	[-1.897, 1.899]
Badminton	-0.037 [-0.083, -0.007]	0.108	-0.012 [-0.037, 0.010]	[-1.546, 1.542]
Table tennis	0.010 [-0.009, 0.043]	0.146	-0.001 [-0.017, 0.017]	[-1.332, 1.328]

Table 18: Factorial interaction contrast at seed zero. Preference uses frames detected in all four arms. Command success uses the full fixed command support. Missingness bounds combine the full-support arm bounds, and are not confidence intervals.

Sport	Preference [95% CI]	Common coverage	Matched success Δ	Missingness bounds
Soccer	0.036 [-0.043, 0.100]	0.191	-0.054 [-0.150, 0.040]	[-2.438, 2.446]
Basketball	n/a	0.000	-0.007 [-0.078, 0.043]	[-3.797, 3.794]
Badminton	0.032 [-0.040, 0.125]	0.108	-0.036 [-0.087, 0.007]	[-3.085, 3.091]
Table tennis	0.007 [-0.032, 0.082]	0.146	-0.019 [-0.056, 0.014]	[-2.664, 2.657]

Table 19: Native minus reduced resolution, Original 2B. Preference uses common observed frames, while success and coverage retain the fixed command support.

Sport	Preference Δ [95% CI]	Common coverage	Matched success Δ	Joint coverage Δ
Soccer	0.023 [-0.005, 0.063]	0.234	0.047 [-0.041, 0.130]	-0.035 [-0.147, 0.063]
Basketball	0.000 [n/a]	0.004	0.044 [0.000, 0.100]	0.019 [-0.092, 0.122]
Badminton	0.002 [-0.039, 0.040]	0.227	0.160 [0.088, 0.242]	0.269 [0.181, 0.351]
Table tennis	0.005 [-0.013, 0.027]	0.226	0.110 [0.062, 0.162]	0.164 [0.082, 0.245]

Table 20: Native minus reduced resolution, Original 14B. Preference uses common observed frames, while success and coverage retain the fixed command support.

Sport	Preference Δ [95% CI]	Common coverage	Matched success Δ	Joint coverage Δ
Soccer	0.030 [-0.008, 0.082]	0.189	0.200 [0.107, 0.297]	0.230 [0.091, 0.370]
Basketball	-0.001 [n/a]	0.009	0.021 [-0.004, 0.049]	-0.003 [-0.075, 0.059]
Badminton	-0.002 [-0.004, 0.000]	0.039	0.199 [0.120, 0.293]	0.317 [0.199, 0.439]
Table tennis	-0.010 [-0.080, 0.057]	0.126	0.123 [0.074, 0.177]	0.303 [0.203, 0.394]

Table 21: Native minus reduced resolution, 2B grid / sport labels. Preference uses common observed frames, while success and coverage retain the fixed command support.

Sport	Preference Δ [95% CI]	Common coverage	Matched success Δ	Joint coverage Δ
Soccer	-0.027 [-0.062, -0.001]	0.196	0.054 [-0.040, 0.142]	0.059 [-0.052, 0.171]
Basketball	0.004 [n/a]	0.007	0.012 [-0.017, 0.047]	-0.003 [-0.056, 0.045]
Badminton	0.012 [-0.001, 0.028]	0.111	0.114 [0.043, 0.204]	0.138 [0.034, 0.259]
Table tennis	0.002 [-0.015, 0.016]	0.191	0.055 [0.009, 0.099]	0.114 [0.005, 0.227]

Table 22: Native minus reduced resolution, 14B grid / sport labels. Preference uses common observed frames, while success and coverage retain the fixed command support.

Sport	Preference Δ [95% CI]	Common coverage	Matched success Δ	Joint coverage Δ
Soccer	-0.015 [-0.099, 0.047]	0.247	0.038 [-0.043, 0.111]	-0.057 [-0.153, 0.040]
Basketball	-0.002 [n/a]	0.004	0.013 [0.002, 0.027]	0.026 [-0.023, 0.078]
Badminton	-0.001 [-0.006, 0.003]	0.104	0.081 [0.034, 0.141]	0.069 [-0.043, 0.176]
Table tennis	-0.040 [-0.113, 0.011]	0.165	0.066 [0.028, 0.111]	0.008 [-0.104, 0.108]

Table 23: Generated seed one minus seed zero, 2B grid / sport labels. Preference uses common observed frames, while success and coverage retain the fixed command support.

Sport	Preference Δ [95% CI]	Common coverage	Matched success Δ	Joint coverage Δ
Soccer	-0.086 [-0.201, -0.010]	0.192	-0.011 [-0.067, 0.049]	0.003 [-0.088, 0.091]
Basketball	n/a	0.000	-0.018 [-0.047, 0.000]	0.000 [-0.068, 0.070]
Badminton	-0.018 [-0.054, 0.000]	0.091	0.028 [-0.008, 0.074]	-0.032 [-0.096, 0.037]
Table tennis	-0.006 [-0.025, 0.015]	0.130	0.026 [-0.022, 0.068]	-0.024 [-0.122, 0.073]

Table 24: Generated seed one minus seed zero, 14B grid / sport labels. Preference uses common observed frames, while success and coverage retain the fixed command support.

Sport	Preference Δ [95% CI]	Common coverage	Matched success Δ	Joint coverage Δ
Soccer	-0.011 [-0.061, 0.030]	0.265	-0.003 [-0.059, 0.049]	-0.002 [-0.121, 0.117]
Basketball	n/a	0.000	-0.004 [-0.009, 0.000]	-0.017 [-0.052, 0.011]
Badminton	-0.005 [-0.023, 0.013]	0.118	0.005 [-0.024, 0.038]	0.005 [-0.061, 0.076]
Table tennis	-0.015 [-0.048, 0.011]	0.144	0.004 [-0.008, 0.019]	-0.045 [-0.132, 0.035]

Table 25: Matched minus removed object channels, 2B grid / sport labels, seed zero. Positive command-success differences favor retaining the object channels. LPIPS differences describe proximity to the recorded future.

Sport	Success Δ [95% CI]	Coverage Δ	LPIPS Δ	Motion Δ
Soccer	-0.005 [-0.016, 0.005]	0.000 [-0.020, 0.019]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]
Basketball	0.002 [-0.003, 0.009]	-0.050 [-0.113, -0.007]	0.000 [0.000, 0.001]	0.000 [0.000, 0.000]
Badminton	-0.002 [-0.011, 0.004]	-0.011 [-0.042, 0.020]	0.000 [-0.001, 0.000]	0.000 [0.000, 0.000]
Table tennis	0.004 [-0.006, 0.018]	0.028 [0.014, 0.043]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]

Table 26: Matched minus removed object channels, 14B grid / sport labels, seed zero. Positive command-success differences favor retaining the object channels. LPIPS differences describe proximity to the recorded future.

Sport	Success Δ [95% CI]	Coverage Δ	LPIPS Δ	Motion Δ
Soccer	-0.001 [-0.011, 0.008]	0.000 [-0.011, 0.012]	0.000 [0.000, 0.002]	0.000 [0.000, 0.000]
Basketball	-0.002 [-0.006, 0.000]	-0.007 [-0.019, 0.004]	-0.001 [-0.001, 0.000]	0.000 [0.000, 0.000]
Badminton	-0.005 [-0.015, 0.004]	0.002 [-0.019, 0.027]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]
Table tennis	0.002 [-0.004, 0.009]	0.022 [-0.005, 0.051]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]

A.5 Robustness within the sports population

Our experiments show limited response to a changed object path (Section 5) despite improved similarity to the recorded future (Section 6). We next ask whether these findings persist across adaptation recipes, resolutions, and control inputs. We retain the same four-sport population and paired-path test, examining robustness within this population rather than transfer to unseen sports. Appendix A provides the complete results.

Adaptation recipes. To extend the matched training comparison beyond one starting checkpoint, we repeat the sport-specific annotation and explicit-grid condition with the second adaptation recipe. Both receive the same 400 clips and 1,000 updates before evaluation on all 200 held-out windows. The 2B and 14B recipes also differ in training history and adapter configuration, so the comparison does not isolate model size.

Neither recipe shows a reliably positive path response in any sport when results are averaged over two generated seeds (Appendix Table 4). Missing-observation bounds contain zero throughout, leaving undetected frames unresolved. A second sampling seed produces no reliable positive change on commonly detected frames, with insufficient observations for the basketball seed contrast.

Resolution and control inputs. To test whether small objects benefit from finer image detail, we increase generation from 256×448 to 512×896 pixels for both original and continued recipes, retaining the same contexts, commands, and seed. Higher resolution improves matched-path proximity in several settings, particularly badminton and table tennis (Figure 7, bottom). Yet no estimable contrast shows a reliably positive improvement in path response (top), and basketball observations remain too sparse for intervals. Resolution also changes detection coverage, so improved proximity cannot be attributed solely to generation.

We also remove the object mask and trajectory channels from both continued recipes at reduced resolution. Matched-path success changes by less than 0.005 in every sport, with all intervals including zero, and mean LPIPS changes by less than 0.0006.

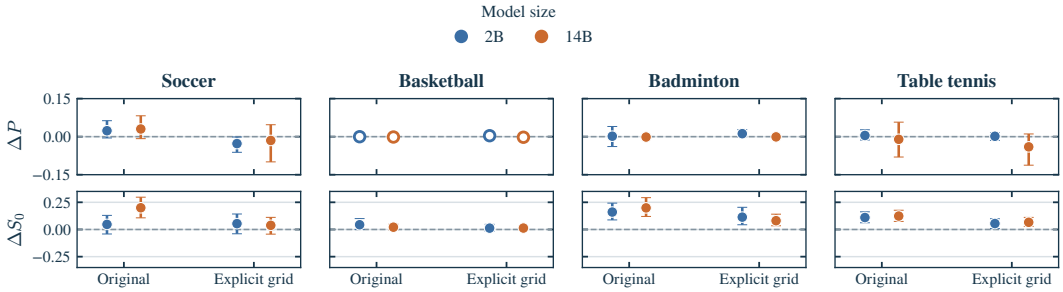

Figure 7: **Higher resolution improves proximity more consistently than path response.** Native-minus-reduced resolution changes in preference ΔP (top) and matched-path success ΔS_0 (bottom), using the same clips and sampling seed. Preference contrasts use frames detected in both pairs, while success counts missing detections as failures. Whiskers show 95% intervals and open markers denote unavailable intervals.

A.6 Detector-confidence sensitivity

Table 27: Confidence sensitivity for Original 2B, render height 256. Threshold changes observation coverage but not the command denominator. Frame counts include both generated seeds where present.

Sport	Threshold	P [95% CI]	Joint coverage	Paired frames	Command frames
Soccer	0.5	-0.002 [-0.016, 0.008]	0.390	190	463
Soccer	0.7	-0.004 [-0.011, 0.000]	0.187	97	463
Soccer	0.9	0.000 [0.000, 0.000]	0.017	11	463
Basketball	0.5	0.001 [0.000, 0.002]	0.079	19	338
Basketball	0.7	0.000 [n/a]	0.028	8	338
Basketball	0.9	n/a	0.000	0	338
Badminton	0.5	0.001 [-0.008, 0.009]	0.320	203	668
Badminton	0.7	-0.001 [-0.008, 0.004]	0.220	133	668
Badminton	0.9	-0.018 [n/a]	0.017	15	668
Table tennis	0.5	-0.017 [-0.037, -0.002]	0.359	277	762
Table tennis	0.7	-0.005 [-0.014, 0.000]	0.240	176	762
Table tennis	0.9	0.000 [-0.001, 0.000]	0.037	31	762

Table 28: Confidence sensitivity for Original 2B, render height 512. Threshold changes observation coverage but not the command denominator. Frame counts include both generated seeds where present.

Sport	Threshold	P [95% CI]	Joint coverage	Paired frames	Command frames
Soccer	0.5	0.015 [-0.004, 0.047]	0.355	219	463
Soccer	0.7	0.010 [0.000, 0.032]	0.176	117	463
Soccer	0.9	0.000 [0.000, 0.000]	0.014	9	463
Basketball	0.5	0.000 [-0.001, 0.000]	0.098	48	338
Basketball	0.7	0.000 [0.000, 0.001]	0.069	34	338
Basketball	0.9	n/a	0.000	0	338
Badminton	0.5	-0.004 [-0.024, 0.015]	0.589	398	668
Badminton	0.7	-0.009 [-0.025, 0.002]	0.502	343	668
Badminton	0.9	0.000 [0.000, 0.000]	0.078	58	668
Table tennis	0.5	0.005 [-0.006, 0.023]	0.523	424	762
Table tennis	0.7	0.003 [-0.004, 0.011]	0.386	316	762
Table tennis	0.9	0.000 [0.000, 0.000]	0.112	100	762

Table 29: Confidence sensitivity for Original 14B, render height 256. Threshold changes observation coverage but not the command denominator. Frame counts include both generated seeds where present.

Sport	Threshold	P [95% CI]	Joint coverage	Paired frames	Command frames
Soccer	0.5	-0.023 [-0.047, -0.003]	0.347	163	463
Soccer	0.7	-0.025 [-0.085, 0.001]	0.111	67	463
Soccer	0.9	0.000 [n/a]	0.021	13	463
Basketball	0.5	0.000 [-0.033, 0.032]	0.080	30	338
Basketball	0.7	0.000 [n/a]	0.019	6	338
Basketball	0.9	n/a	0.000	0	338
Badminton	0.5	-0.008 [-0.036, 0.009]	0.100	73	668
Badminton	0.7	0.000 [0.000, 0.001]	0.041	32	668
Badminton	0.9	0.000 [n/a]	0.005	4	668
Table tennis	0.5	-0.023 [-0.058, 0.004]	0.221	165	762
Table tennis	0.7	0.003 [-0.009, 0.022]	0.089	68	762
Table tennis	0.9	0.000 [0.000, 0.000]	0.018	15	762

Table 30: Confidence sensitivity for Original 14B, render height 512. Threshold changes observation coverage but not the command denominator. Frame counts include both generated seeds where present.

Sport	Threshold	P [95% CI]	Joint coverage	Paired frames	Command frames
Soccer	0.5	0.000 [-0.009, 0.007]	0.577	278	463
Soccer	0.7	-0.003 [-0.011, 0.003]	0.310	136	463
Soccer	0.9	0.000 [0.000, 0.000]	0.026	8	463
Basketball	0.5	0.000 [-0.001, 0.000]	0.076	39	338
Basketball	0.7	0.000 [-0.001, 0.000]	0.056	30	338
Basketball	0.9	n/a	0.000	0	338
Badminton	0.5	0.001 [-0.002, 0.005]	0.417	302	668
Badminton	0.7	0.001 [-0.001, 0.006]	0.332	244	668
Badminton	0.9	-0.003 [-0.010, 0.000]	0.059	50	668
Table tennis	0.5	0.000 [-0.023, 0.030]	0.524	423	762
Table tennis	0.7	-0.002 [-0.006, 0.001]	0.385	319	762
Table tennis	0.9	0.000 [0.000, 0.000]	0.117	103	762

Table 31: Confidence sensitivity for 2B grid / sport labels, render height 256. Threshold changes observation coverage but not the command denominator. Frame counts include both generated seeds where present.

Sport	Threshold	P [95% CI]	Joint coverage	Paired frames	Command frames
Soccer	0.5	0.002 [-0.005, 0.009]	0.336	351	926
Soccer	0.7	0.002 [-0.001, 0.007]	0.142	153	926
Soccer	0.9	-0.001 [n/a]	0.006	5	926
Basketball	0.5	0.000 [-0.001, 0.000]	0.045	27	676
Basketball	0.7	0.000 [0.000, 0.000]	0.028	14	676
Basketball	0.9	n/a	0.000	0	676
Badminton	0.5	-0.002 [-0.012, 0.007]	0.184	248	1336
Badminton	0.7	-0.010 [-0.031, 0.000]	0.108	145	1336
Badminton	0.9	0.000 [n/a]	0.008	13	1336
Table tennis	0.5	-0.003 [-0.009, 0.002]	0.320	503	1524
Table tennis	0.7	-0.004 [-0.011, 0.000]	0.203	317	1524
Table tennis	0.9	0.000 [0.000, 0.000]	0.031	50	1524

Table 32: Confidence sensitivity for 2B grid / sport labels, render height 512. Threshold changes observation coverage but not the command denominator. Frame counts include both generated seeds where present.

Sport	Threshold	P [95% CI]	Joint coverage	Paired frames	Command frames
Soccer	0.5	0.001 [-0.004, 0.007]	0.394	208	463
Soccer	0.7	-0.001 [-0.002, 0.000]	0.177	91	463
Soccer	0.9	-0.001 [n/a]	0.009	6	463
Basketball	0.5	0.000 [0.000, 0.001]	0.042	17	338
Basketball	0.7	0.000 [n/a]	0.024	10	338
Basketball	0.9	n/a	0.000	0	338
Badminton	0.5	0.003 [-0.002, 0.010]	0.338	234	668
Badminton	0.7	0.002 [-0.007, 0.011]	0.263	183	668
Badminton	0.9	0.019 [0.000, 0.052]	0.042	35	668
Table tennis	0.5	-0.006 [-0.014, 0.000]	0.446	357	762
Table tennis	0.7	-0.006 [-0.021, 0.002]	0.301	254	762
Table tennis	0.9	0.000 [0.000, 0.000]	0.116	100	762

Table 33: Confidence sensitivity for 2B grid / original labels, render height 256. Threshold changes observation coverage but not the command denominator. Frame counts include both generated seeds where present.

Sport	Threshold	P [95% CI]	Joint coverage	Paired frames	Command frames
Soccer	0.5	0.013 [-0.007, 0.052]	0.433	206	463
Soccer	0.7	0.045 [0.000, 0.147]	0.228	114	463
Soccer	0.9	0.000 [0.000, 0.000]	0.025	13	463
Basketball	0.5	0.000 [-0.001, 0.002]	0.031	7	338
Basketball	0.7	0.000 [n/a]	0.020	3	338
Basketball	0.9	n/a	0.000	0	338
Badminton	0.5	0.001 [-0.008, 0.012]	0.269	185	668
Badminton	0.7	0.000 [0.000, 0.000]	0.162	109	668
Badminton	0.9	0.000 [0.000, 0.001]	0.020	13	668
Table tennis	0.5	-0.004 [-0.013, 0.004]	0.320	245	762
Table tennis	0.7	0.002 [0.000, 0.005]	0.180	135	762
Table tennis	0.9	0.000 [0.000, 0.001]	0.034	31	762

Table 34: Confidence sensitivity for 2B flat / sport labels, render height 256. Threshold changes observation coverage but not the command denominator. Frame counts include both generated seeds where present.

Sport	Threshold	P [95% CI]	Joint coverage	Paired frames	Command frames
Soccer	0.5	-0.006 [-0.024, 0.011]	0.365	189	463
Soccer	0.7	0.000 [-0.001, 0.000]	0.196	91	463
Soccer	0.9	0.000 [n/a]	0.010	4	463
Basketball	0.5	0.000 [0.000, 0.000]	0.068	24	338
Basketball	0.7	0.000 [n/a]	0.018	5	338
Basketball	0.9	n/a	0.000	0	338
Badminton	0.5	0.000 [0.000, 0.001]	0.230	144	668
Badminton	0.7	0.000 [0.000, 0.001]	0.122	76	668
Badminton	0.9	0.002 [n/a]	0.013	9	668
Table tennis	0.5	0.002 [-0.003, 0.011]	0.348	248	762
Table tennis	0.7	0.000 [0.000, 0.000]	0.231	165	762
Table tennis	0.9	0.000 [0.000, 0.001]	0.026	20	762

Table 35: Confidence sensitivity for 2B flat / original labels, render height 256. Threshold changes observation coverage but not the command denominator. Frame counts include both generated seeds where present.

Sport	Threshold	P [95% CI]	Joint coverage	Paired frames	Command frames
Soccer	0.5	0.003 [-0.007, 0.014]	0.426	211	463
Soccer	0.7	0.005 [0.000, 0.011]	0.208	109	463
Soccer	0.9	0.000 [-0.001, 0.000]	0.026	14	463
Basketball	0.5	-0.014 [-0.049, 0.001]	0.061	25	338
Basketball	0.7	0.000 [n/a]	0.007	4	338
Basketball	0.9	n/a	0.000	0	338
Badminton	0.5	0.017 [0.000, 0.037]	0.214	143	668
Badminton	0.7	0.000 [0.000, 0.000]	0.110	70	668
Badminton	0.9	0.000 [0.000, 0.000]	0.008	6	668
Table tennis	0.5	-0.003 [-0.005, 0.000]	0.341	263	762
Table tennis	0.7	0.002 [-0.002, 0.006]	0.205	155	762
Table tennis	0.9	0.000 [0.000, 0.001]	0.020	17	762

Table 36: Confidence sensitivity for 14B grid / sport labels, render height 256. Threshold changes observation coverage but not the command denominator. Frame counts include both generated seeds where present.

Sport	Threshold	P [95% CI]	Joint coverage	Paired frames	Command frames
Soccer	0.5	-0.017 [-0.037, -0.001]	0.431	452	926
Soccer	0.7	-0.001 [-0.005, 0.001]	0.223	242	926
Soccer	0.9	0.000 [0.000, 0.000]	0.018	21	926
Basketball	0.5	0.001 [-0.001, 0.004]	0.015	14	676
Basketball	0.7	0.000 [n/a]	0.007	6	676
Basketball	0.9	n/a	0.000	0	676
Badminton	0.5	-0.006 [-0.016, 0.000]	0.221	294	1336
Badminton	0.7	0.001 [0.000, 0.004]	0.132	187	1336
Badminton	0.9	0.007 [n/a]	0.014	25	1336
Table tennis	0.5	-0.005 [-0.016, 0.006]	0.346	535	1524
Table tennis	0.7	-0.012 [-0.025, 0.000]	0.202	324	1524
Table tennis	0.9	0.000 [0.000, 0.000]	0.041	63	1524

Table 37: Confidence sensitivity for 14B grid / sport labels, render height 512. Threshold changes observation coverage but not the command denominator. Frame counts include both generated seeds where present.

Sport	Threshold	P [95% CI]	Joint coverage	Paired frames	Command frames
Soccer	0.5	-0.021 [-0.068, 0.000]	0.375	190	463
Soccer	0.7	0.000 [-0.002, 0.000]	0.168	78	463
Soccer	0.9	0.000 [n/a]	0.030	5	463
Basketball	0.5	0.000 [0.000, 0.001]	0.049	22	338
Basketball	0.7	0.000 [n/a]	0.020	8	338
Basketball	0.9	n/a	0.000	0	338
Badminton	0.5	0.001 [0.000, 0.002]	0.288	200	668
Badminton	0.7	0.001 [0.000, 0.002]	0.232	160	668
Badminton	0.9	-0.003 [-0.009, 0.000]	0.023	20	668
Table tennis	0.5	-0.008 [-0.042, 0.018]	0.377	302	762
Table tennis	0.7	0.012 [0.000, 0.024]	0.244	203	762
Table tennis	0.9	0.000 [0.000, 0.000]	0.079	70	762

Table 38: **Basketball detector-threshold sensitivity.** The fixed 0.5 threshold remains primary. The 0.3 sensitivity reruns the detector on the same seed-zero videos. P is conditional preference, C is joint coverage, and intervals are 95%. At threshold 0.3, unchanged real videos give $P = 0.00$, while known translations give $P = 0.98$.

Training condition	$P_{0.5}$	$C_{0.5}$	$P_{0.3}$	$C_{0.3}$
2B flat / original labels	-0.014 [-0.049, 0.001]	0.06	0.005 [0.000, 0.015]	0.15
2B flat / sport labels	0.000 [0.000, 0.000]	0.07	-0.013 [-0.103, 0.060]	0.15
2B grid / original labels	0.000 [-0.001, 0.002]	0.03	-0.002 [-0.009, 0.003]	0.10
2B grid / sport labels	0.000 [-0.001, 0.000]	0.05	0.065 [0.001, 0.157]	0.10

Lowering the basketball threshold increases support in every training condition. The sport-specific explicit-grid arm is positive at this exploratory threshold, but the factorial intersection contains only two clips and cannot support an interval. The threshold was evaluated after observing sparse primary support, so it does not replace the fixed primary analysis or establish a factor-level improvement.

A.7 Secondary trajectory displacement

Table 39: Original 2B, render height 256. Secondary trajectory response uses the same observed frames as primary preference. R_x is horizontal displacement divided by the signed command shift. V and D are absolute vertical change and total displacement in normalized coordinates. Gain measures movement, not assigned-path accuracy, and need not lie in $[-1, 1]$. Generated seeds are averaged within clips where two are available.

Sport	R_x [95% CI]	V	D	Joint coverage	n
Soccer	-0.012 [-0.088, 0.054]	0.009	0.014	0.390	28
Basketball	0.000 [-0.001, 0.001]	0.001	0.001	0.079	7
Badminton	0.002 [-0.027, 0.030]	0.003	0.006	0.320	33
Table tennis	-0.032 [-0.123, 0.027]	0.008	0.019	0.359	39

Table 40: Original 2B, render height 512. Secondary trajectory response uses the same observed frames as primary preference. R_x is horizontal displacement divided by the signed command shift. V and D are absolute vertical change and total displacement in normalized coordinates. Gain measures movement, not assigned-path accuracy, and need not lie in $[-1, 1]$. Generated seeds are averaged within clips where two are available.

Sport	R_x [95% CI]	V	D	Joint coverage	n
Soccer	0.064 [-0.040, 0.220]	0.007	0.017	0.355	24
Basketball	0.000 [-0.002, 0.001]	0.001	0.001	0.098	7
Badminton	-0.022 [-0.116, 0.068]	0.013	0.024	0.589	38
Table tennis	0.016 [-0.019, 0.072]	0.008	0.012	0.523	46

Table 41: Original 14B, render height 256. Secondary trajectory response uses the same observed frames as primary preference. R_x is horizontal displacement divided by the signed command shift. V and D are absolute vertical change and total displacement in normalized coordinates. Gain measures movement, not assigned-path accuracy, and need not lie in $[-1, 1]$. Generated seeds are averaged within clips where two are available.

Sport	R_x [95% CI]	V	D	Joint coverage	n
Soccer	-0.133 [-0.237, -0.043]	0.023	0.038	0.347	32
Basketball	0.004 [-0.152, 0.159]	0.017	0.025	0.080	13
Badminton	-0.069 [-0.215, 0.049]	0.015	0.028	0.100	20
Table tennis	-0.083 [-0.234, 0.021]	0.035	0.049	0.221	45

Table 42: Original 14B, render height 512. Secondary trajectory response uses the same observed frames as primary preference. R_x is horizontal displacement divided by the signed command shift. V and D are absolute vertical change and total displacement in normalized coordinates. Gain measures movement, not assigned-path accuracy, and need not lie in $[-1, 1]$. Generated seeds are averaged within clips where two are available.

Sport	R_x [95% CI]	V	D	Joint coverage	n
Soccer	-0.037 [-0.108, 0.010]	0.013	0.021	0.577	36
Basketball	0.000 [-0.001, 0.001]	0.001	0.001	0.076	10
Badminton	0.003 [-0.003, 0.013]	0.002	0.003	0.417	33
Table tennis	-0.003 [-0.063, 0.083]	0.008	0.015	0.524	43

Table 43: 2B grid / sport labels, render height 256. Secondary trajectory response uses the same observed frames as primary preference. R_x is horizontal displacement divided by the signed command shift. V and D are absolute vertical change and total displacement in normalized coordinates. Gain measures movement, not assigned-path accuracy, and need not lie in $[-1, 1]$. Generated seeds are averaged within clips where two are available.

Sport	R_x [95% CI]	V	D	Joint coverage	n
Soccer	0.012 [-0.031, 0.059]	0.011	0.022	0.336	30
Basketball	-0.002 [-0.003, 0.000]	0.001	0.001	0.045	9
Badminton	-0.009 [-0.041, 0.009]	0.015	0.017	0.184	32
Table tennis	-0.018 [-0.040, 0.002]	0.007	0.011	0.320	42

Table 44: 2B grid / sport labels, render height 512. Secondary trajectory response uses the same observed frames as primary preference. R_x is horizontal displacement divided by the signed command shift. V and D are absolute vertical change and total displacement in normalized coordinates. Gain measures movement, not assigned-path accuracy, and need not lie in $[-1, 1]$. Generated seeds are averaged within clips where two are available.

Sport	R_x [95% CI]	V	D	Joint coverage	n
Soccer	0.005 [-0.026, 0.043]	0.004	0.010	0.394	29
Basketball	0.001 [-0.003, 0.004]	0.001	0.001	0.042	7
Badminton	0.009 [-0.010, 0.036]	0.003	0.004	0.338	33
Table tennis	-0.010 [-0.036, 0.012]	0.005	0.008	0.446	43

Table 45: 2B grid / original labels, render height 256. Secondary trajectory response uses the same observed frames as primary preference. R_x is horizontal displacement divided by the signed command shift. V and D are absolute vertical change and total displacement in normalized coordinates. Gain measures movement, not assigned-path accuracy, and need not lie in $[-1, 1]$. Generated seeds are averaged within clips where two are available.

Sport	R_x [95% CI]	V	D	Joint coverage	n
Soccer	0.025 [-0.053, 0.146]	0.010	0.021	0.433	30
Basketball	-0.003 [-0.008, 0.001]	0.001	0.002	0.031	5
Badminton	0.009 [-0.067, 0.099]	0.015	0.020	0.269	30
Table tennis	-0.027 [-0.102, 0.019]	0.008	0.016	0.320	40

Table 46: 2B flat / sport labels, render height 256. Secondary trajectory response uses the same observed frames as primary preference. R_x is horizontal displacement divided by the signed command shift. V and D are absolute vertical change and total displacement in normalized coordinates. Gain measures movement, not assigned-path accuracy, and need not lie in $[-1, 1]$. Generated seeds are averaged within clips where two are available.

Sport	R_x [95% CI]	V	D	Joint coverage	n
Soccer	-0.077 [-0.192, 0.001]	0.009	0.024	0.365	26
Basketball	0.000 [-0.001, 0.002]	0.000	0.001	0.068	8
Badminton	0.001 [0.000, 0.002]	0.001	0.001	0.230	30
Table tennis	0.018 [-0.013, 0.065]	0.001	0.006	0.348	41

Table 47: 2B flat / original labels, render height 256. Secondary trajectory response uses the same observed frames as primary preference. R_x is horizontal displacement divided by the signed command shift. V and D are absolute vertical change and total displacement in normalized coordinates. Gain measures movement, not assigned-path accuracy, and need not lie in $[-1, 1]$. Generated seeds are averaged within clips where two are available.

Sport	R_x [95% CI]	V	D	Joint coverage	n
Soccer	0.002 [-0.028, 0.028]	0.006	0.016	0.426	30
Basketball	-0.018 [-0.068, 0.006]	0.001	0.004	0.061	8
Badminton	0.066 [-0.001, 0.148]	0.014	0.021	0.214	28
Table tennis	-0.011 [-0.027, 0.003]	0.012	0.016	0.341	41

Table 48: 14B grid / sport labels, render height 256. Secondary trajectory response uses the same observed frames as primary preference. R_x is horizontal displacement divided by the signed command shift. V and D are absolute vertical change and total displacement in normalized coordinates. Gain measures movement, not assigned-path accuracy, and need not lie in $[-1, 1]$. Generated seeds are averaged within clips where two are available.

Sport	R_x [95% CI]	V	D	Joint coverage	n
Soccer	-0.142 [-0.272, -0.029]	0.016	0.042	0.431	34
Basketball	-0.003 [-0.007, 0.000]	0.001	0.001	0.015	5
Badminton	-0.037 [-0.072, -0.005]	0.009	0.013	0.221	32
Table tennis	-0.036 [-0.117, 0.024]	0.013	0.020	0.346	45

Table 49: 14B grid / sport labels, render height 512. Secondary trajectory response uses the same observed frames as primary preference. R_x is horizontal displacement divided by the signed command shift. V and D are absolute vertical change and total displacement in normalized coordinates. Gain measures movement, not assigned-path accuracy, and need not lie in $[-1, 1]$. Generated seeds are averaged within clips where two are available.

Sport	R_x [95% CI]	V	D	Joint coverage	n
Soccer	-0.034 [-0.119, 0.020]	0.015	0.021	0.375	31
Basketball	0.001 [0.000, 0.003]	0.001	0.001	0.049	7
Badminton	0.002 [-0.001, 0.009]	0.001	0.003	0.288	30
Table tennis	-0.008 [-0.075, 0.042]	0.008	0.017	0.377	40

Table 50: Known native image translation. Secondary trajectory response uses the same observed frames as primary preference. R_x is horizontal displacement divided by the signed command shift. V and D are absolute vertical change and total displacement in normalized coordinates. Gain measures movement, not assigned-path accuracy, and need not lie in $[-1, 1]$. Generated seeds are averaged within clips where two are available.

Sport	R_x [95% CI]	V	D	Joint coverage	n
Soccer	1.007 [0.936, 1.076]	0.008	0.162	0.898	39
Basketball	0.993 [0.978, 1.000]	0.004	0.150	0.903	36
Badminton	0.956 [0.856, 1.011]	0.009	0.157	0.971	42
Table tennis	0.953 [0.897, 0.999]	0.015	0.160	0.905	47

Table 51: Known image translation after resampling. Secondary trajectory response uses the same observed frames as primary preference. R_x is horizontal displacement divided by the signed command shift. V and D are absolute vertical change and total displacement in normalized coordinates. Gain measures movement, not assigned-path accuracy, and need not lie in $[-1, 1]$. Generated seeds are averaged within clips where two are available.

Sport	R_x [95% CI]	V	D	Joint coverage	n
Soccer	0.851 [0.733, 0.941]	0.019	0.173	0.667	34
Basketball	0.995 [0.975, 1.004]	0.006	0.151	0.759	31
Badminton	0.975 [0.905, 1.022]	0.009	0.158	0.932	42
Table tennis	0.933 [0.878, 0.977]	0.016	0.162	0.849	46

Table 52: Unchanged native image. Secondary trajectory response uses the same observed frames as primary preference. R_x is horizontal displacement divided by the signed command shift. V and D are absolute vertical change and total displacement in normalized coordinates. Gain measures movement, not assigned-path accuracy, and need not lie in $[-1, 1]$. Generated seeds are averaged within clips where two are available.

Sport	R_x [95% CI]	V	D	Joint coverage	n
Soccer	0.000 [0.000, 0.000]	0.000	0.000	1.000	40
Basketball	0.000 [0.000, 0.000]	0.000	0.000	1.000	37
Badminton	0.000 [0.000, 0.000]	0.000	0.000	1.000	42
Table tennis	0.000 [0.000, 0.000]	0.000	0.000	1.000	48

Table 53: Unchanged resampled image. Secondary trajectory response uses the same observed frames as primary preference. R_x is horizontal displacement divided by the signed command shift. V and D are absolute vertical change and total displacement in normalized coordinates. Gain measures movement, not assigned-path accuracy, and need not lie in $[-1, 1]$. Generated seeds are averaged within clips where two are available.

Sport	R_x [95% CI]	V	D	Joint coverage	n
Soccer	0.000 [0.000, 0.000]	0.000	0.000	0.704	35
Basketball	0.000 [0.000, 0.000]	0.000	0.000	0.813	33
Badminton	0.000 [0.000, 0.000]	0.000	0.000	0.969	42
Table tennis	0.000 [0.000, 0.000]	0.000	0.000	0.925	48

A.8 LOCALIZATION DETAIL, REFERENCE PROXIMITY, AND MOTION

Table 54: Localization detail for Original 2B, render height 256. Distances are observed Euclidean errors in normalized image coordinates. Success fractions use all command frames. Subscripts identify matched and shifted arms.

Sport	d_0	d_1	$S_0(0.02)$	$S_1(0.02)$	$S_0(0.05)$	$S_1(0.05)$
Soccer	0.265	0.309	0.134	0.000	0.165	0.010
Basketball	0.145	0.196	0.005	0.000	0.010	0.000
Badminton	0.192	0.249	0.060	0.000	0.111	0.001
Table tennis	0.226	0.262	0.044	0.000	0.071	0.007

Table 55: Localization detail for Original 2B, render height 512. Distances are observed Euclidean errors in normalized image coordinates. Success fractions use all command frames. Subscripts identify matched and shifted arms.

Sport	d_0	d_1	$S_0(0.02)$	$S_1(0.02)$	$S_0(0.05)$	$S_1(0.05)$
Soccer	0.234	0.324	0.152	0.000	0.211	0.014
Basketball	0.108	0.216	0.028	0.000	0.054	0.000
Badminton	0.118	0.221	0.170	0.000	0.271	0.000
Table tennis	0.146	0.216	0.103	0.001	0.181	0.006

Table 56: Localization detail for Original 14B, render height 256. Distances are observed Euclidean errors in normalized image coordinates. Success fractions use all command frames. Subscripts identify matched and shifted arms.

Sport	d_0	d_1	$S_0(0.02)$	$S_1(0.02)$	$S_0(0.05)$	$S_1(0.05)$
Soccer	0.364	0.406	0.046	0.000	0.056	0.006
Basketball	0.315	0.437	0.004	0.000	0.011	0.003
Badminton	0.293	0.373	0.006	0.000	0.008	0.000
Table tennis	0.384	0.387	0.010	0.000	0.023	0.009

Table 57: Localization detail for Original 14B, render height 512. Distances are observed Euclidean errors in normalized image coordinates. Success fractions use all command frames. Subscripts identify matched and shifted arms.

Sport	d_0	d_1	$S_0(0.02)$	$S_1(0.02)$	$S_0(0.05)$	$S_1(0.05)$
Soccer	0.234	0.301	0.188	0.000	0.256	0.006
Basketball	0.093	0.227	0.018	0.000	0.032	0.000
Badminton	0.123	0.206	0.146	0.002	0.207	0.007
Table tennis	0.136	0.199	0.073	0.002	0.146	0.008

Table 58: Localization detail for 2B grid / sport labels, render height 256. Distances are observed Euclidean errors in normalized image coordinates. Success fractions use all command frames. Subscripts identify matched and shifted arms.

Sport	d_0	d_1	$S_0(0.02)$	$S_1(0.02)$	$S_0(0.05)$	$S_1(0.05)$
Soccer	0.304	0.355	0.082	0.001	0.125	0.015
Basketball	0.269	0.304	0.005	0.000	0.009	0.004
Badminton	0.239	0.266	0.054	0.000	0.075	0.001
Table tennis	0.219	0.249	0.038	0.003	0.077	0.009

Table 59: Localization detail for 2B grid / sport labels, render height 512. Distances are observed Euclidean errors in normalized image coordinates. Success fractions use all command frames. Subscripts identify matched and shifted arms.

Sport	d_0	d_1	$S_0(0.02)$	$S_1(0.02)$	$S_0(0.05)$	$S_1(0.05)$
Soccer	0.246	0.312	0.129	0.000	0.184	0.007
Basketball	0.077	0.211	0.022	0.000	0.030	0.000
Badminton	0.108	0.199	0.119	0.000	0.174	0.001
Table tennis	0.160	0.220	0.070	0.002	0.119	0.005

Table 60: Localization detail for 2B grid / original labels, render height 256. Distances are observed Euclidean errors in normalized image coordinates. Success fractions use all command frames. Subscripts identify matched and shifted arms.

Sport	d_0	d_1	$S_0(0.02)$	$S_1(0.02)$	$S_0(0.05)$	$S_1(0.05)$
Soccer	0.256	0.309	0.142	0.001	0.175	0.008
Basketball	0.191	0.305	0.018	0.000	0.020	0.000
Badminton	0.223	0.288	0.058	0.000	0.092	0.002
Table tennis	0.202	0.254	0.034	0.002	0.067	0.007

Table 61: Localization detail for 2B flat / sport labels, render height 256. Distances are observed Euclidean errors in normalized image coordinates. Success fractions use all command frames. Subscripts identify matched and shifted arms.

Sport	d_0	d_1	$S_0(0.02)$	$S_1(0.02)$	$S_0(0.05)$	$S_1(0.05)$
Soccer	0.232	0.296	0.133	0.001	0.176	0.016
Basketball	0.295	0.312	0.035	0.000	0.040	0.002
Badminton	0.203	0.264	0.062	0.000	0.090	0.001
Table tennis	0.200	0.243	0.045	0.000	0.075	0.002

Table 62: Localization detail for 2B flat / original labels, render height 256. Distances are observed Euclidean errors in normalized image coordinates. Success fractions use all command frames. Subscripts identify matched and shifted arms.

Sport	d_0	d_1	$S_0(0.02)$	$S_1(0.02)$	$S_0(0.05)$	$S_1(0.05)$
Soccer	0.303	0.367	0.111	0.001	0.167	0.015
Basketball	0.150	0.269	0.029	0.000	0.035	0.000
Badminton	0.228	0.283	0.050	0.002	0.085	0.003
Table tennis	0.211	0.238	0.026	0.004	0.058	0.008

Table 63: Localization detail for 14B grid / sport labels, render height 256. Distances are observed Euclidean errors in normalized image coordinates. Success fractions use all command frames. Subscripts identify matched and shifted arms.

Sport	d_0	d_1	$S_0(0.02)$	$S_1(0.02)$	$S_0(0.05)$	$S_1(0.05)$
Soccer	0.299	0.354	0.127	0.002	0.164	0.011
Basketball	0.316	0.465	0.002	0.000	0.002	0.000
Badminton	0.209	0.263	0.051	0.000	0.078	0.000
Table tennis	0.215	0.263	0.037	0.002	0.069	0.007

Table 64: Localization detail for 14B grid / sport labels, render height 512. Distances are observed Euclidean errors in normalized image coordinates. Success fractions use all command frames. Subscripts identify matched and shifted arms.

Sport	d_0	d_1	$S_0(0.02)$	$S_1(0.02)$	$S_0(0.05)$	$S_1(0.05)$
Soccer	0.200	0.279	0.146	0.000	0.204	0.012
Basketball	0.111	0.223	0.012	0.000	0.017	0.000
Badminton	0.097	0.192	0.122	0.001	0.157	0.008
Table tennis	0.134	0.193	0.080	0.000	0.133	0.010

Table 65: Reference-proximity and motion diagnostics for Original 2B, 256 by 448, matched command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.293 [0.276, 0.312]	0.071	0.037	0.417
Basketball	0.322 [0.300, 0.348]	0.096	0.039	0.099
Badminton	0.223 [0.190, 0.254]	0.053	0.016	0.350
Table tennis	0.195 [0.161, 0.233]	0.049	0.014	0.385

Table 66: Reference-proximity and motion diagnostics for Original 2B, 256 by 448, shifted command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.293 [0.276, 0.312]	0.071	0.037	0.431
Basketball	0.322 [0.300, 0.348]	0.096	0.039	0.086
Badminton	0.223 [0.190, 0.254]	0.053	0.016	0.351
Table tennis	0.195 [0.161, 0.233]	0.049	0.014	0.391

Table 67: Reference-proximity and motion diagnostics for Original 2B, 512 by 896, matched command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.192 [0.174, 0.213]	0.066	0.037	0.393
Basketball	0.210 [0.185, 0.240]	0.085	0.039	0.109
Badminton	0.143 [0.103, 0.182]	0.042	0.015	0.604
Table tennis	0.136 [0.097, 0.182]	0.042	0.012	0.561

Table 68: Reference-proximity and motion diagnostics for Original 2B, 512 by 896, shifted command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.192 [0.174, 0.213]	0.066	0.037	0.376
Basketball	0.210 [0.185, 0.240]	0.085	0.039	0.113
Badminton	0.143 [0.103, 0.182]	0.042	0.015	0.620
Table tennis	0.136 [0.097, 0.181]	0.042	0.012	0.572

Table 69: Reference-proximity and motion diagnostics for Original 14B, 256 by 448, matched command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.582 [0.560, 0.608]	0.140	0.071	0.415
Basketball	0.606 [0.584, 0.629]	0.177	0.075	0.136
Badminton	0.573 [0.551, 0.601]	0.161	0.071	0.126
Table tennis	0.466 [0.419, 0.516]	0.129	0.045	0.251

Table 70: Reference-proximity and motion diagnostics for Original 14B, 256 by 448, shifted command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.582 [0.560, 0.608]	0.139	0.071	0.378
Basketball	0.606 [0.584, 0.629]	0.177	0.075	0.134
Badminton	0.573 [0.550, 0.601]	0.161	0.070	0.138
Table tennis	0.466 [0.419, 0.516]	0.129	0.045	0.279

Table 71: Reference-proximity and motion diagnostics for Original 14B, 512 by 896, matched command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.208 [0.189, 0.228]	0.065	0.040	0.592
Basketball	0.214 [0.189, 0.244]	0.085	0.039	0.076
Badminton	0.146 [0.105, 0.184]	0.042	0.016	0.447
Table tennis	0.142 [0.103, 0.189]	0.045	0.013	0.556

Table 72: Reference-proximity and motion diagnostics for Original 14B, 512 by 896, shifted command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.208 [0.189, 0.229]	0.065	0.040	0.594
Basketball	0.214 [0.189, 0.243]	0.085	0.039	0.090
Badminton	0.146 [0.105, 0.183]	0.042	0.016	0.428
Table tennis	0.142 [0.103, 0.189]	0.045	0.013	0.559

Table 73: Reference-proximity and motion diagnostics for 2B grid / sport labels, 256 by 448, matched command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.309 [0.292, 0.331]	0.073	0.034	0.356
Basketball	0.343 [0.319, 0.369]	0.102	0.040	0.047
Badminton	0.224 [0.192, 0.255]	0.054	0.016	0.228
Table tennis	0.203 [0.168, 0.242]	0.051	0.014	0.349

Table 74: Reference-proximity and motion diagnostics for 2B grid / sport labels, 256 by 448, shifted command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.309 [0.292, 0.330]	0.073	0.034	0.357
Basketball	0.343 [0.319, 0.369]	0.102	0.040	0.061
Badminton	0.224 [0.192, 0.255]	0.054	0.016	0.210
Table tennis	0.203 [0.168, 0.242]	0.051	0.014	0.353

Table 75: Reference-proximity and motion diagnostics for 2B grid / sport labels, 256 by 448, removed command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.311 [0.292, 0.333]	0.072	0.035	0.342
Basketball	0.347 [0.321, 0.374]	0.103	0.042	0.097
Badminton	0.225 [0.193, 0.256]	0.054	0.015	0.241
Table tennis	0.203 [0.168, 0.242]	0.052	0.014	0.336

Table 76: Reference-proximity and motion diagnostics for 2B grid / sport labels, 512 by 896, matched command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.217 [0.196, 0.239]	0.068	0.034	0.414
Basketball	0.230 [0.202, 0.260]	0.088	0.040	0.042
Badminton	0.145 [0.104, 0.182]	0.042	0.014	0.348
Table tennis	0.139 [0.099, 0.187]	0.043	0.012	0.478

Table 77: Reference-proximity and motion diagnostics for 2B grid / sport labels, 512 by 896, shifted command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.217 [0.196, 0.239]	0.068	0.034	0.410
Basketball	0.230 [0.202, 0.260]	0.088	0.040	0.050
Badminton	0.145 [0.104, 0.182]	0.042	0.014	0.361
Table tennis	0.139 [0.099, 0.187]	0.043	0.012	0.478

Table 78: Reference-proximity and motion diagnostics for 2B grid / original labels, 256 by 448, matched command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.306 [0.288, 0.325]	0.071	0.038	0.471
Basketball	0.343 [0.319, 0.369]	0.102	0.045	0.042
Badminton	0.226 [0.194, 0.257]	0.054	0.017	0.308
Table tennis	0.204 [0.169, 0.242]	0.052	0.015	0.350

Table 79: Reference-proximity and motion diagnostics for 2B grid / original labels, 256 by 448, shifted command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.306 [0.288, 0.325]	0.071	0.038	0.444
Basketball	0.343 [0.319, 0.369]	0.102	0.045	0.046
Badminton	0.226 [0.194, 0.258]	0.054	0.017	0.298
Table tennis	0.204 [0.169, 0.242]	0.052	0.015	0.357

Table 80: Reference-proximity and motion diagnostics for 2B flat / sport labels, 256 by 448, matched command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.306 [0.289, 0.326]	0.071	0.038	0.394
Basketball	0.340 [0.316, 0.365]	0.101	0.045	0.110
Badminton	0.223 [0.192, 0.254]	0.054	0.016	0.270
Table tennis	0.202 [0.167, 0.240]	0.051	0.014	0.376

Table 81: Reference-proximity and motion diagnostics for 2B flat / sport labels, 256 by 448, shifted command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.306 [0.289, 0.326]	0.071	0.038	0.389
Basketball	0.340 [0.316, 0.365]	0.101	0.045	0.075
Badminton	0.223 [0.192, 0.254]	0.054	0.016	0.270
Table tennis	0.202 [0.167, 0.240]	0.051	0.014	0.371

Table 82: Reference-proximity and motion diagnostics for 2B flat / original labels, 256 by 448, matched command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.313 [0.294, 0.333]	0.073	0.037	0.439
Basketball	0.344 [0.319, 0.370]	0.102	0.043	0.067
Badminton	0.228 [0.196, 0.259]	0.055	0.017	0.236
Table tennis	0.206 [0.171, 0.245]	0.052	0.015	0.384

Table 83: Reference-proximity and motion diagnostics for 2B flat / original labels, 256 by 448, shifted command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.313 [0.294, 0.333]	0.073	0.037	0.442
Basketball	0.344 [0.319, 0.370]	0.102	0.043	0.128
Badminton	0.228 [0.196, 0.259]	0.055	0.017	0.248
Table tennis	0.206 [0.171, 0.245]	0.052	0.015	0.384

Table 84: Reference-proximity and motion diagnostics for 14B grid / sport labels, 256 by 448, matched command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.315 [0.298, 0.334]	0.071	0.039	0.444
Basketball	0.342 [0.321, 0.365]	0.102	0.045	0.020
Badminton	0.226 [0.194, 0.256]	0.054	0.018	0.261
Table tennis	0.205 [0.172, 0.244]	0.052	0.014	0.386

Table 85: Reference-proximity and motion diagnostics for 14B grid / sport labels, 256 by 448, shifted command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.315 [0.298, 0.334]	0.071	0.039	0.451
Basketball	0.343 [0.321, 0.366]	0.102	0.045	0.018
Badminton	0.226 [0.194, 0.256]	0.054	0.018	0.247
Table tennis	0.205 [0.172, 0.244]	0.052	0.014	0.383

Table 86: Reference-proximity and motion diagnostics for 14B grid / sport labels, 256 by 448, removed command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.319 [0.300, 0.340]	0.071	0.040	0.446
Basketball	0.350 [0.326, 0.377]	0.102	0.047	0.040
Badminton	0.229 [0.195, 0.261]	0.054	0.018	0.246
Table tennis	0.204 [0.170, 0.243]	0.052	0.015	0.376

Table 87: Reference-proximity and motion diagnostics for 14B grid / sport labels, 512 by 896, matched command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.206 [0.186, 0.230]	0.063	0.038	0.400
Basketball	0.230 [0.204, 0.259]	0.086	0.041	0.049
Badminton	0.144 [0.104, 0.180]	0.040	0.015	0.308
Table tennis	0.138 [0.099, 0.184]	0.042	0.011	0.401

Table 88: Reference-proximity and motion diagnostics for 14B grid / sport labels, 512 by 896, shifted command. LPIPS and pixel error compare with the one recorded future and are not visual-quality scores. Motion includes camera motion.

Sport	LPIPS [95% CI]	Pixel error	Temporal change	Command coverage
Soccer	0.206 [0.186, 0.230]	0.063	0.038	0.410
Basketball	0.229 [0.204, 0.259]	0.086	0.041	0.084
Badminton	0.144 [0.104, 0.180]	0.040	0.015	0.301
Table tennis	0.138 [0.099, 0.184]	0.042	0.011	0.394

Table 89: Frozen-last-context-frame reference. Its generated temporal change is exactly zero. Reference motion is measured on the recorded future.

Sport	LPIPS [95% CI]	Pixel error	Generated change	Real change
Soccer	0.285 [0.263, 0.307]	0.082	0.000	0.045
Basketball	0.294 [0.258, 0.331]	0.107	0.000	0.045
Badminton	0.156 [0.110, 0.199]	0.039	0.000	0.018
Table tennis	0.157 [0.113, 0.202]	0.046	0.000	0.015

Table 90: Mean clip-level absolute pixel difference between matched and intervened outputs, normalized to [0,1], at sampling seed zero. Pixel response does not establish object adherence. Fixed seeds also retain the inherited numerical nondeterminism described in the engineering protocol. Source-bootstrap intervals remain in the numerical report.

Recipe	Height	Intervention	Soccer	Basketball	Badminton	Table tennis
Original 2B	256	shifted	0.006	0.007	0.006	0.006
Original 2B	512	shifted	0.007	0.007	0.005	0.006
Original 14B	256	shifted	0.012	0.011	0.011	0.009
Original 14B	512	shifted	0.008	0.010	0.005	0.006
2B grid / sport labels	256	shifted	0.008	0.007	0.007	0.005
2B grid / sport labels	256	removed	0.008	0.011	0.006	0.006
2B grid / sport labels	512	shifted	0.007	0.008	0.006	0.005
2B grid / original labels	256	shifted	0.007	0.009	0.007	0.006
2B flat / sport labels	256	shifted	0.007	0.009	0.006	0.006
2B flat / original labels	256	shifted	0.008	0.009	0.006	0.006
14B grid / sport labels	256	shifted	0.008	0.010	0.006	0.006
14B grid / sport labels	256	removed	0.008	0.011	0.008	0.006
14B grid / sport labels	512	shifted	0.007	0.008	0.006	0.005

A.9 Observational mismatches and engineering-item exclusion

Table 91: Secondary observational mismatch diagnostic for matched outputs. Entries are other-clip command distance minus own-command distance on the same detected frames, with source-bootstrap intervals. Positive values favor the own command but do not establish a causal response, since source contexts differ. The full report retains command-overlap counts, coverage, and both component distances for every row.

Recipe (render height)	Soccer	Basketball	Badminton	Table tennis
Original 2B (256)	0.219 [0.119, 0.318]	0.030 [-0.169, 0.181]	0.094 [0.030, 0.163]	0.031 [-0.010, 0.073]
Original 2B (512)	0.265 [0.181, 0.352]	0.214 [n/a]	0.139 [0.087, 0.195]	0.102 [0.071, 0.134]
Original 14B (256)	0.065 [-0.026, 0.164]	-0.110 [-0.228, 0.013]	0.033 [-0.089, 0.129]	0.025 [-0.030, 0.081]
Original 14B (512)	0.223 [0.138, 0.307]	0.107 [n/a]	0.156 [0.080, 0.220]	0.102 [0.067, 0.141]
2B grid / sport labels (256)	0.193 [0.140, 0.261]	0.090 [n/a]	0.057 [-0.039, 0.144]	0.037 [0.000, 0.076]
2B grid / sport labels (512)	0.222 [0.105, 0.332]	0.069 [n/a]	0.164 [0.090, 0.235]	0.101 [0.053, 0.154]
2B grid / original labels (256)	0.230 [0.131, 0.332]	0.255 [n/a]	0.048 [-0.043, 0.139]	0.064 [0.030, 0.101]
2B flat / sport labels (256)	0.227 [0.124, 0.329]	-0.028 [-0.233, 0.162]	0.082 [-0.059, 0.201]	0.055 [0.027, 0.084]
2B flat / original labels (256)	0.189 [0.102, 0.282]	0.139 [n/a]	0.084 [0.006, 0.170]	0.048 [0.016, 0.088]
14B grid / sport labels (256)	0.142 [0.032, 0.248]	0.074 [n/a]	0.083 [0.004, 0.162]	0.081 [0.048, 0.119]
14B grid / sport labels (512)	0.200 [0.106, 0.296]	0.187 [n/a]	0.137 [0.065, 0.198]	0.116 [0.073, 0.166]

Table 92: Sensitivity omitting the four clips used in engineering pilots, one per sport. The primary reporting population remains all 200 clips. This table averages two generated seeds within each remaining clip at reduced resolution.

Recipe	Sport	Clips	P [95% CI]	Joint coverage	Matched success
2B grid / sport labels	Soccer	49	0.002 [-0.005, 0.009]	0.329	0.128
2B grid / sport labels	Basketball	49	0.000 [-0.001, 0.000]	0.045	0.009
2B grid / sport labels	Badminton	49	-0.002 [-0.012, 0.007]	0.188	0.077
2B grid / sport labels	Table tennis	49	-0.003 [-0.009, 0.002]	0.327	0.079
14B grid / sport labels	Soccer	49	-0.018 [-0.038, 0.000]	0.429	0.168
14B grid / sport labels	Basketball	49	0.001 [-0.001, 0.004]	0.015	0.002
14B grid / sport labels	Badminton	49	-0.006 [-0.016, 0.000]	0.227	0.080
14B grid / sport labels	Table tennis	49	-0.005 [-0.016, 0.006]	0.353	0.070

B Metric Interpretation and Sampling Units

A bounded preference statistic. For any point y and two commands a, b, the reverse triangle inequality gives $-\|a-b\|_2 \le \|y-b\|_2 - \|y-a\|_2 \le \|a-b\|_2$. Applying this bound to the two terms in Eq. 1 gives $P \in [-1, 1]$. Setting $y^0 = y^1$ makes their sum exactly zero. Setting $y^0 = c^0$ and $y^1 = c^1$ gives one. These properties isolate preference under the paired intervention, but do not make preference an exact-distance score. For example, collinear predictions beyond their assigned commands can also achieve one. This is why observed distance and command-proximity success accompany preference throughout the report.

Movement can differ from preference and accuracy. As a secondary diagnostic, we define horizontal response gain

$$R_{it,x} = \frac{y^1_{it,x} - y^0_{it,x}}{c^1_{it,x} - c^0_{it,x}}. \tag{3}$$

It uses exactly the jointly observed frames of the primary preference statistic. A gain of one indicates movement by the commanded horizontal displacement, and zero indicates unchanged horizontal position. Neither establishes accurate localization. For example, commands at horizontal positions 0.20 and 0.35 and outputs at 0.60 and 0.75, all with the same vertical coordinate, give $R_x = 1$ but $P = 0$ and zero proximity success at threshold 0.05. Unlike preference, gain need not lie in $[-1, 1]$. We also report mean absolute vertical response and mean displacement magnitude in normalized coordinates. Missing observations retain no gain estimate, and every gain table includes joint coverage and measured clip counts. The same clip, seed, and source aggregation applies as in the primary analysis. This secondary diagnostic was defined during campaign execution before any complete primary paired report was available and does not replace the fixed preference endpoint.

Frames, clips, seeds, and sources. The frame score is averaged over jointly observed frames within each clip and seed. The two generated seeds, where present, are then averaged within a clip. A conditional average uses the measured seeds when the other seed has no jointly detected frames, and the report retains the seed-specific observation counts. Worst-case bounds retain every commanded frame in both seeds. Clips without any machine command have no command-dependent score and remain in total population counts. Sports with multiple source recordings use a source-cluster bootstrap that includes every selected clip of a sampled recording. The statistic weights clips equally, and source clusters can contain different numbers of selected clips. Soccer uses a clip bootstrap conditional on its single selected match. Intervals are withheld when fewer than five clips have a measured metric or when fewer than 95% of bootstrap resamples contain a measured clip. They are descriptive pointwise intervals, with no correction for multiple comparisons.

Matched contrasts. The annotation effect averages the sport-specific-minus-original annotation difference over the two alignment choices. The alignment effect averages the explicit-grid-minus-flattened difference over the two annotation choices. Their interaction is the difference between those annotation effects at the two alignment choices. Preference contrasts use seed zero and the intersection of detected frames across all four arms. Resolution contrasts similarly intersect observed frames across both resolution pairs. These conditional contrasts do not describe frames outside the

intersection. The corresponding missingness bounds combine each arm's lower or upper bound according to the sign of its contrast coefficient. Command-proximity and coverage contrasts use the same fixed command support and retain missing detections as unsuccessful localization measurements.

C Detector Protocol and Model-Assisted Semantic Inspection

Detector execution. We use the pinned WASB implementation at revision `923462cacdeb3353b84ddebdedb3f4b7a8553b0f`. Soccer, basketball, and badminton weights come from the WASB project. The table-tennis weights are its WASB baseline distributed by the BlurBall project. Checkpoint, architecture, and configuration hashes accompany every detection cache. Input frames are RGB, transformed by a centered affine warp to 288×512, and normalized using ImageNet channel statistics. The network observes three real consecutive frames and predicts three heat maps. Windows advance by three frames, with a final overlapping window when needed. We select the connected component with greatest heat-map mass above threshold 0.5 and use its intensity-weighted centroid. The detector receives no command coordinates and does not interpolate a missing observation. Thresholds 0.7 and 0.9 are sensitivity analyses on the saved confidence values. For basketball, an additional threshold-0.3 sensitivity reruns heat-map postprocessing on the same seed-zero videos and repeats the unchanged and known-translation real-video controls.

Known transformations. The positive image-space control translates each real future by its item's signed horizontal displacement of 134 pixels using a black image border. The negative control supplies the same video to both arms while keeping the two commands distinct. The resampling control reduces real frames to 256×448 and expands them back to 512×896 using the same area interpolation and integer conversion as the generator's output path. Translated and resampled arrays are scored directly, without an additional video encode. This isolates transformation response from an extra compression pass, while generated outputs are decoded from their saved H.264 files. Because native commands originate from the same detector, perfect native matched-command proximity is definitional and is not presented as independent semantic accuracy.

Blinded model-assisted point inspection. Eight items per sport were selected by a fixed hash independently of detector confidence. For each item, the coding assistant inspected a four-frame sheet and attempted to localize the object in a designated frame before seeing the detector's prediction. The labels, unresolved decisions, and coordinate convention were saved before computing point errors. The reviewer was an AI model, not a human annotator. This small inspection is a limited diagnostic and cannot validate performance on the full population or replace independent human annotation. The 10-pixel and 20-pixel error summaries are descriptive thresholds. Unresolved points are excluded from coordinate-error summaries and remain visible in the counts.

Table 93: Model-assisted point inspection, not human ground truth. Resolved means the assistant identified a visible object. Detected and within-10-pixel counts apply only to those resolved cases. Unresolved and no-visible-object cases remain explicit.

Sport	Reviewed	Resolved	Unresolved	Not visible	Detected	≤ 10 px
Soccer	8	8	0	0	3	2
Basketball	8	6	2	0	2	2
Badminton	8	5	3	0	4	3
Table tennis	8	3	4	1	3	2

The inspection includes confidently visible objects that were not detected, large false localizations, close-ups, and broadcast transitions. These cases support the explicit limitations of the measurement pipeline. They do not justify silently filtering difficult clips or treating detector-derived training labels as semantically verified.

D Implementation Audit and Training Provenance

Temporal output correction. The original evaluation wrapper requested the desired future length from a pipeline that includes reconstructed context in its returned video. This placed context frames in the future-scoring window. An end-to-end comparison against real frame indices reproduced the offset. The corrected wrapper requests enough frames for context plus future, rounds up to the causal VAE's valid length, and then extracts the requested future interval. For five context frames and 21 future frames, it requests 29 frames and returns frames 5 through 25. Regression tests cover several context lengths and both fourfold and eightfold causal temporal compression. All 5,600 campaign videos use the corrected future-only protocol. Earlier engineering pilots and the original 174-item audit remain separate and are not pooled with these outputs.

Spatial and temporal alignment. The original adapter resizes a flattened token sequence when its feature-grid dimensions differ from the transformer grid. This operation need not preserve image coordinates, and the training path used a different spatial target from the sampling path. The explicit-grid path uses the transformer's actual VAE and patch dimensions. It inserts zero features over the context and trailing padding, retains the first causal frame separately, and averages each subsequent four-frame temporal group. Spatial resizing then targets the two patch-grid dimensions before flattening. Coordinate-field tests verify spatial preservation, temporal grouping, padding, and rejection of incompatible lengths.

Frozen representations and gradients. The VAE and text encoder remain frozen. The training cache uses float32 pixel normalization, contiguous channel-first layout, the pretrained reciprocal latent standard deviations, and bfloat16 VAE execution. On three actual training clips, cached latents were bitwise equal to the live encoding path. A separate objective test verifies equivalence of cached and live latent inputs. Every continuation run checks for nonzero adapter gradients on its first step. Losses, gradient norms, item order, optimizer state, and random states are retained. The complete training configuration uses AdamW with betas $(0.9, 0.999)$, weight decay 0.01, learning rate 10^{-4}, batch size one, and gradient clipping at norm one. There is no adherence-based early stopping.

Execution optimizations. Text embeddings are cached for the fixed sport prompts using the pretrained encoder. Cache validation includes encoder and tokenizer hashes, software versions, and the embedding implementation. The text encoder is omitted from GPU placement when these cached embeddings are used. This avoids an unnecessary transient allocation without changing the sampled embeddings. The inherited execution protocol uses bfloat16 and cuDNN kernel benchmarking. Fixed sampling seeds therefore do not guarantee bitwise-identical pixels across repeated processes or devices. Engineering repeats and the memory-placement comparison are retained separately from the primary study. Generation runs in fixed index shards that are merged only after the exact item and seed population is complete. The merged report retains each shard's configuration and index hashes.

Table 94: Training provenance. Starting step counts come from the existing adapted checkpoints and are not a controlled pretraining budget. Times cover the optimization loop, including checkpoint writes and shared-device contention, but exclude model and prompt-cache setup. M denotes millions of trainable parameters.

Continuation	Start step	Rank	Stride	Updates	Params M	Minutes
2B flat / original labels	138000	32	1	1000	178.7	21.7
2B flat / sport labels	138000	32	1	1000	178.7	31.6
2B grid / original labels	138000	32	1	1000	178.7	24.6
2B grid / sport labels	138000	32	1	1000	178.7	23.2
14B grid / sport labels	72000	16	3	1000	423.2	40.3

E Original-Control Exploratory Audit

Before the fixed campaign, an exploratory audit evaluated 174 existing items using their original commanded trajectories and independent sport-specific detection. Real footage did not separate

reliably from within-sport mismatched commands under that original annotation pipeline. The original heuristic object extractor was designed around soccer imagery and could select graphics or unrelated visual elements in the other sports. This observation motivated the replacement machine annotations and fresh generations. It is evidence about the original evaluation inputs, not proof that every adapted generator ignores object controls. The archived exploratory report retains its own commands, outputs, and temporal protocol. Its scores are not compared numerically with the corrected paired intervention as if they estimated the same quantity.

Table 95: Exploratory original-command audit on real footage. This historical statistic is $1 - d_{\text{own}}/d_{\text{mismatch}}$ and differs from the new paired-intervention preference. Intervals use the original item bootstrap. None excludes zero.

Sport	Items	Score [95% CI]	Coverage	Measured items
Soccer	50	0.053 [-0.101, 0.189]	0.292	25
Basketball	28	0.014 [-0.516, 0.363]	0.185	10
Badminton	50	0.066 [-0.133, 0.219]	0.831	46
Table tennis	46	-0.045 [-0.209, 0.092]	0.771	41

F Reproduction and Artifact Organization

The campaign is defined by the fixed benchmark manifests, the experiment protocol and its documented development amendment, and the generation-plan file. The amendment followed engineering pilots and preceded primary generated-output scoring. Four selected evaluation clips, one per sport, were also used in the engineering pilots. The primary results retain the fixed population, and Table 92 reports the corrected recipes with those four clips omitted. The selected population and update budgets did not change in response to primary adherence results. The reproduction sequence prepares the sport-specific detectors, selects and validates evaluation and training windows, caches frozen representations, executes five continuation runs, generates all 24 rows, scores completed rows, and aggregates the paired analyses. The paper report generator reads the completed numerical report and produces its tables and plots. The main paper build does not read the retired synthetic placeholder tables.

Input and model hashes identify the exact locally processed artifacts. Durable archives preserve the original manifests and a path-relocation map for restoring the execution layout. Source media and pretrained weights retain their respective permissions. The local archive is not a public release, and the numerical study does not imply that its source footage may be redistributed.

G Real-Video Figure Frame Sources

Figure 1(a,b) uses a new illustrative paired experiment on close-up clip `swc_soccer_2dfbbf17b9d6`. The clip was selected for a large visible ball, independently of the new generated outcomes. We visually checked agent-assisted ball-centre annotations in all 21 recorded future frames, with approximate five-pixel precision. These annotations are not independent human ground truth. Command A follows those coordinates. Command B uses $x_B(t) = x_A(0) - 1.5[x_A(t) - x_A(0)]$ and $y_B(t) = y_A(t)$, giving a 440-pixel final horizontal separation. Only the object-mask and object-trajectory control channels change. All other controls, input frames, prompt, checkpoint, and inference settings remain fixed within each pair. We generated three paired seeds (0, 1, 2), using the corrected 2B checkpoint, 20 inference steps, guidance 1.0, and native 512×896 resolution. Seed zero and the endpoint display time were fixed before generation. The images use the same crop, $x \in [40, 680)$ and $y \in [300, 475)$, without moving or resizing individual objects. White circles are visual highlights, not quantitative localization results. This divergent-command illustration is separate from the fixed-translation benchmark. Its full-clip LPIPS values for A/B are 0.4477/0.4475, 0.4420/0.4427, and 0.4480/0.4478 for seeds 0, 1, and 2, respectively, versus 0.4266 for repeating the last input frame. Thus this close-up illustrates weak intervention response, not a future-similarity improvement. Panel (c) instead summarizes the unchanged 200-clip, four-sport benchmark with equal clip weights. The local provenance file `figure1_spatial_commands.json` records source hashes, frame indices, crops,

commands, and displayed measurements. The scripts `prepare_figure1_experiment.py`, `score_figure1_experiment.py`, and `design_figure1.py` reproduce the preparation, scoring, and rendering. These benchmark frames retain the source permissions described above.

Figure 2 uses real frames from the fixed evaluation population. The four sport examples are selected to show recognizable playing surfaces, independently of generated outcomes. The paired-command example uses badminton clip `swc_badminton_c331a4c95798`, selected for continuous, mostly vertical command geometry that makes the horizontal translation legible. Both overlaid paths retain all 21 command positions and their exact 134-pixel separation on the native canvas. The shared background is the last input frame. The paths are machine annotations, not independent human ground truth. Panel (c) specifies the two comparisons without displaying experimental outcomes. The local provenance file `sportworldbench_design.json` records all five source clips, frame indices, crop coordinates, commands, and hashes. The script `plot_sportworldbench_design.py` reproduces the figure from the archived benchmark inputs. The script `design_wrapped_figures.py` reflows both figure masters to three-inch width, preserving their source images, path coordinates, and measurements. These frames retain the benchmark source permissions described above.

H Extended Limitations and Responsible Use

What is measured. In the sports experiments, the detector both supplies replacement machine annotations and measures generated object positions. Misses and false localizations can therefore affect the result, and adaptation toward detector-specific artifacts could improve a score without improving the true ball trajectory. Known-translation calibration checks response, while the small model-assisted visual audit is not independent human ground truth. Missingness bounds address unobserved frames but do not correct incorrect detections. Our strongest claims concern measured command response under this evaluation pipeline.

What generalizes. The experiments cover four sports and a smaller extension to animals and road vehicles, with short futures and two existing adaptation recipes. The non-sports extension tests frozen sports-adapted checkpoints without new-domain training and cannot establish the performance attainable after such training. Soccer contains only one held-out match. Source separation is limited to available adaptation-training records and identified recordings. A single continuation-training seed and a fixed 1,000-update budget do not characterize all training outcomes or convergence. The two model recipes differ in several factors besides size. Null or negative results therefore cannot establish an intrinsic limitation of diffusion models.

What is not tested. A translated image-space command need not be physically feasible given the other controls. Preference for that command is not evidence of correct sport rules, physical dynamics, player identity, or long-horizon interaction. Changing render resolution also changes context detail, latent-grid size, and the detector's effective view of small objects. Reference similarity to one recorded future does not measure the full distribution of plausible futures. No human preference study was conducted.

Data and generated footage. Sports broadcasts depict identifiable people and real events. Generated continuations should be identified as synthetic. Redistributing footage or weights remains subject to applicable permissions, consent, and license terms recorded in the experiment artifacts.

I Non-Sports Domain Transfer

Population and source separation. We construct the non-sports extension from the DAVIS 2017 train/validation release (Pont-Tuset et al., 2017). Animal sources are `bear`, `blackswan`, `cows`, `dog`, `elephant`, `flamingo`, and `sheep`. Road-vehicle sources are `bus`, `car-roundabout`, `car-shadow`, `car-turn`, `scooter-black`, and `scooter-gray`. Source names and first-frame inspection exclude sporting, racing, and stunt scenes, including the ambiguous `classic-car` event. We take up to two nonoverlapping 26-frame windows beginning at frames 0 and 26, retaining only complete windows. This yields 13 animal windows from seven sources and

10 road-vehicle windows from six sources. Selection precedes examination of generated outputs and retains every selected window during scoring. No DAVIS adaptation is performed, although exposure during backbone pretraining is unknown.

Target paths and generation. DAVIS supplies instance masks and category labels. Within each source, the target is the largest annotated instance of its specified category in the first frame, and its published instance identity is retained throughout. The center of the instance-mask bounding box defines each future path point. The paired instruction shifts these centers horizontally by 134 pixels, with direction fixed by an item hash. Both paths use the same in-frame annotation support. Each window contains five context frames and 21 future frames, with consecutive source frames played at 25 frames per second without temporal interpolation. We resize to 512×896 and use the existing object-center disk and Gaussian trajectory channels. Sports-specific channels, their availability flags, and the sport-identity vector are zero. The text prompt describes the target category in an outdoor or road scene. The two final sport-specific, explicit-grid checkpoints remain fixed, with two generated seeds, 20 sampling steps, and guidance scale 1.0. The resulting grid contains 184 videos, separate from the 5,600-video sports campaign.

Independent localization and calibration. We use Faster R-CNN (Ren et al., 2015) with torchvision's ResNet50-FPN v2 COCO_V1 weights. At fixed confidence 0.5, each frame contributes the center of its highest-confidence target-category box. The detector receives neither the instructed path nor the recorded future when scoring generation. This category-level selection can switch instances, so high detection coverage does not establish correct instance association. We evaluate real-video localization against the published annotation centers, as well as unchanged and known-translation response controls, before scoring generated outputs. Localization success uses normalized distance 0.05 and counts missing detections as failures. Missingness bounds cover absent observations but do not correct wrong-instance detections.

Reporting and interpretation. We retain the sports study's preference, coverage, proximity, missingness, and LPIPS definitions. Two seeds are averaged within each window before 2,000 bootstrap draws over source videos, separately by domain and model recipe. Intervals are descriptive and pointwise. The first bear window also serves as an engineering generation check and remains in the reporting population. Its independently repeated matched-path output is identical at the decoded-pixel level for seed zero, a single-case check rather than a population repeatability study. Differences in categories, object sizes, control availability, annotation sources, and detectors mean that comparison with sports does not isolate a causal effect of domain. The extension tests transfer of fixed sports-adapted checkpoints, not the performance attainable after non-sports adaptation. The public-source metadata declares CC BY-NC 4.0, and the artifacts preserve dataset, checkpoint, command, and output hashes.

Table 96: **Non-sports real-video calibration.** Published DAVIS instance-mask boxes supply independent reference centers. Translation uses the same fixed horizontal intervention as generation. Localization success uses distance 0.05, counting missing detections as failures. Intervals resample source videos. Unchanged-video preference is zero to numerical precision in both domains.

Domain	Windows	Sources	P translated	Real coverage	Real localization success
Animals	13	7	0.834 [0.642, 0.946]	0.996 [0.989, 1.000]	0.864 [0.619, 0.997]
Road vehicles	10	6	0.818 [0.715, 0.923]	1.000 [1.000, 1.000]	0.895 [0.792, 0.989]

Table 97: **Non-sports path response.** Fixed sports-adapted checkpoints, native resolution, two generated seeds per window. C is joint coverage and S_0, S_1 are matched- and shifted-path success at distance 0.05. Missingness bounds are identification bounds rather than confidence intervals. No further training is performed on DAVIS.

Domain	Recipe	P (95% CI)	C	S_0	S_1	Missingness bounds
Animals	2B	-0.0018 [-0.0091, 0.0038]	0.987	0.540	0.016	[-0.015, 0.011]
Animals	14B	-0.0009 [-0.0030, 0.0015]	0.978	0.597	0.013	[-0.023, 0.021]
Road vehicles	2B	-0.0114 [-0.0355, 0.0021]	0.998	0.688	0.000	[-0.013, -0.009]
Road vehicles	14B	0.0021 [-0.0007, 0.0052]	0.983	0.709	0.007	[-0.015, 0.019]

Table 98: **Non-sports future similarity.** Similarity gain is repeated-frame LPIPS minus matched-path LPIPS. Path effect is shifted-path minus matched-path LPIPS. Positive gain favors generation. Pointwise 95% intervals resample source videos after averaging generated seeds within windows.

Domain	Recipe	Matched LPIPS	Similarity gain	Path effect
Animals	2B	0.293 [0.202, 0.376]	0.027 [-0.015, 0.066]	0.00008 [-0.00019, 0.00034]
Animals	14B	0.286 [0.196, 0.372]	0.035 [-0.017, 0.082]	0.00039 [-0.00030, 0.00146]
Road vehicles	2B	0.345 [0.261, 0.420]	0.095 [0.078, 0.114]	0.00023 [-0.00006, 0.00051]
Road vehicles	14B	0.343 [0.266, 0.408]	0.097 [0.073, 0.123]	-0.00049 [-0.00151, 0.00024]