SANITY CHECKS FOR PLANNING WITH VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video world models support planning by predicting the outcomes of candidate plans. We examine how strongly those predictions depend on the plan and how consistently the evaluator selects it. We introduce three checks that compare plan changes with sampler noise, compare predictions with simulator futures under alternative plans, and repeat the evaluator's decision on identical inputs. In World-in-World image-goal navigation with Wan2.2, we analyze 33 episodes and 536 replanning steps. Changing one action affects the goal-similarity score less than changing the sampler seed: the ratio is 0.81 (95% CI $[0.69, 0.95]$) for the VLM planner and 0.57 ($[0.54, 0.60]$) for the heuristic planner. Predictions favor their own plan's true future over perturbed plans' futures by a margin equal to 1.5% of prediction error. At 43% of evaluated steps, none of five repeated VLM queries reproduces the executed plan. Four training-free confidence signals have episode-success AUCs of 0.47–0.66. Prediction fidelity has the opposite association ($\mathrm{AUC} = 0.12$), with more accurate predictions in failed episodes. These measurements distinguish action sensitivity, decision consistency, and task success, and provide a practical protocol for evaluating confidence signals in video-based planners.

1 INTRODUCTION

Video generation models are increasingly used as world models for embodied planning. At each decision point the agent proposes candidate action sequences, imagines the outcome of each with the world model, scores the imagined futures, and executes the best one (Du et al., 2023; Bar et al., 2025; Zhang et al., 2026). The recipe is the video counterpart of model-predictive control with learned dynamics (Chua et al., 2018; Hansen et al., 2024), and closed-loop benchmarks such as World-in-World (Zhang et al., 2026) now measure whether it improves task success.

Deployment raises a further question: when should an agent act on what it imagined? Confidence estimation for video world models is an active topic (Mei et al., 2025), and a recent position paper asks evaluations to measure action fidelity and the value of abstaining on uncertain rollouts (Yu et al., 2026). For an off-the-shelf generator, the available training-free signals are few: the agreement among rollouts drawn with different sampler seeds (Lakshminarayanan et al., 2017; Wang et al., 2023), the sensitivity of the imagined outcome to perturbed actions, likelihood proxies from the denoiser, and the stated confidence of a vision-language evaluator (Tian et al., 2023).

Interpreting these signals requires separating three properties: whether a prediction responds to the plan, whether the evaluator repeats its decision, and whether either property predicts success. A visually accurate rollout can reflect a persistent scene without capturing the consequences of an action. Likewise, agreement across evaluator samples measures decision consistency, which need not imply correctness. Closed-loop success rates alone do not identify these distinctions.

We study them through controlled perturbations, following the evaluation approach of Adebayo et al. (2018).

Check 1 compares how much a change of plan moves the imagined outcome with how much a change of sampler seed moves it. **Check 2** compares the imagination with futures that the simulator renders under the executed plan and under perturbed plans. **Check 3** re-queries the planner's evaluator on identical inputs and counts how often it reproduces its own decision. We run the checks inside

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.

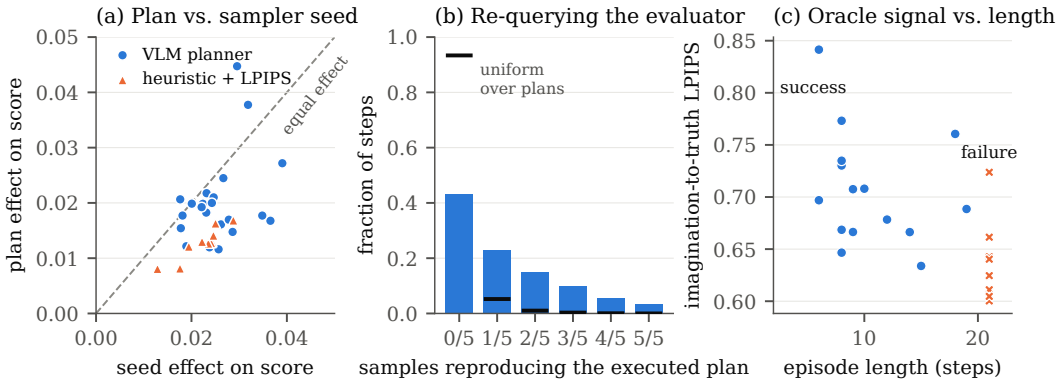


Figure 1: **The three checks.** (a) Per-episode plan effect against seed effect on the goal-similarity score (Eq. 2); points below the dashed line are episodes in which changing one action of the plan moved the score less than changing the seed. (b) Fraction of steps at which $0, \ldots, 5$ of five re-queried evaluator samples reproduce the executed plan (bars), against uniform choice among plans of the executed length (black ticks). (c) Per-episode distance between imagined and true futures against episode length; every failed episode ends at the 21-step budget.

World-in-World's image-goal navigation task with its default world model, Wan2.2-TI2V-5B (Wan Team, 2025), instrumenting 33 episodes and 536 replanning steps under both planners the benchmark provides, and we then ask what training-free confidence signals achieve on the same episodes.

The measurements identify substantial sampler and evaluator variation. Plan changes affect the goal-similarity score less than seed changes under both planners (Figure 1a). The corresponding offline ratio ranges from 0.90 to 0.13 across prompt and resolution settings (Table 2). Predictions only slightly favor their own plan's true future, and evaluator agreement varies across repeated queries. Episode-level confidence analysis then tests whether these quantities predict task success.

Our contributions are:

- A protocol for measuring plan sensitivity relative to sampler noise, plan-specific prediction fidelity, and evaluator agreement.
- An evaluation of the protocol in a published closed-loop benchmark, including both provided planners and four generator configurations.
- An episode-level analysis connecting these measurements to confidence estimation and distinguishing prediction fidelity from task success.

2 Related Work

World models for planning. Learned dynamics models have long supported planning and policy learning (Ha & Schmidhuber, 2018; Chua et al., 2018; Janner et al., 2019; Hafner et al., 2023; Hansen et al., 2024). Generative video models extend the idea to pixels: interactive environments learned from video (Bruce et al., 2024), foundation models for physical AI (NVIDIA et al., 2025), text-guided video policies (Du et al., 2023), and navigation world models trained with trajectory conditioning (Koh et al., 2021; Bar et al., 2025). World-in-World (Zhang et al., 2026) evaluates such models in closed loop and reports that task success, not visual quality (Huang et al., 2024), is the relevant target. We use its harness unchanged.

Measuring action conditioning. Genie quantifies controllability by comparing frames generated from inferred latent actions with frames generated from random ones (Bruce et al., 2024). Hansen & Wang (2026) name the failure we study *action-marginalized hallucination* and detect it by shuffling actions within a batch, in a world model they train; they treat it as a data-coverage problem and reduce it through data collection and fine-tuning. Yu et al. (2026) propose interventional action fidelity, measured by branching matched histories under different actions, as an evaluation protocol. Our

reference scale is sampler variation at a fixed plan. We study a frozen, text-conditioned generator inside a published planning system and relate these measurements to confidence estimation.

Uncertainty for dynamics and video models. Ensemble disagreement gates model use in model-based RL (Chua et al., 2018; Janner et al., 2019; Lakshminarayanan et al., 2017). C^3 learns calibrated dense uncertainty for action-conditioned video generation (Mei et al., 2025), and Berger et al. (2026) show that ensemble disagreement in latent dynamics models can miss compounding error. For language models, agreement across samples (Wang et al., 2023; Kuhn et al., 2023) and verbalized confidence (Tian et al., 2023) are the standard training-free signals. Selective prediction (Geifman & El-Yaniv, 2017), conformal risk control (Angelopoulos et al., 2024), and help-seeking robot planners (Ren et al., 2023) convert such signals into decisions, provided the signals carry information about the outcome. Our experiments assess this relationship at the episode level.

Rigour in evaluation. Adebayo et al. (2018) showed that saliency methods can produce convincing maps that do not depend on the model they explain. Henderson et al. (2018) and Agarwal et al. (2021) showed that evaluation in deep RL depends on how seed variation and uncertainty are measured. We follow their practice of reporting interval estimates over independent units, which here are episodes rather than steps.

3 SETTING

Task. World-in-World image-goal navigation places an agent in a Habitat (Savva et al., 2019) scene from HM3D (Ramakrishnan et al., 2021) and gives it an image taken at the goal. The primitive actions are *forward 0.2 m*, *turn left 22.5°*, and *turn right 22.5°*. Success is judged by the benchmark's distance criterion, within a budget of 21 replanning steps.

World model. The default world model is Wan2.2-TI2V-5B (Wan Team, 2025). The harness converts a plan $a = (a_1, \ldots, a_H)$ into a text prompt with one phrase per primitive, padded to 13 primitives with *forward*, and generates a 33-frame clip at 1280×704 with 50 denoising steps (guidance scale 5) from the current observation. We write $W(o, a, \xi)$ for the imagined video, where ξ is the sampler seed; given o and a, the seed is the only source of randomness.

Planners. The benchmark ships two planners. The *VLM planner* uses Qwen2.5-VL-72B (Bai et al., 2025) to propose candidate plans, shows it their imagined futures, and asks it for the plan to execute. The *heuristic planner* samples candidates and executes the one whose imagined final frame is most similar to the goal image under

$$S(W(o, a, \xi)) = 1 - \mathrm{LPIPS}\big(o_{\mathrm{goal}},\ \text{final frame of } W(o, a, \xi)\big) \tag{1}$$

(Zhang et al., 2018). Both execute the first $\max(H - 2, 1)$ primitives and replan. Plans have 1 to 5 primitives; 78% have 5.

Instrumentation. At every replanning step, after the plan a is chosen and before it is executed, we (i) re-imagine a under $k = 5$ seeds $\xi_0, \ldots, \xi_4$ and imagine three perturbed plans a'_1, a'_2, a'_3, each differing from a in a single primitive, under ξ_0; (ii) checkpoint the simulator, render the true future under a and under each a'_m, and restore the state; and (iii) for the VLM planner, send the evaluator's original request again with five samples. None of this alters what the agent observes or does. We ran 23 episodes with the VLM planner (326 steps) and 10 with the heuristic planner (210 steps). Rendering true futures was enabled for the VLM planner only. Instrumentation costs about four H100 GPU-hours per episode (Appendix F).

Statistics. Steps within an episode are dependent, so every episode-level claim uses an episode-clustered bootstrap (10^4 resamples) or a sign-flip permutation test on per-episode means (Appendix A). Discrimination is the Mann–Whitney AUC with a permutation p-value. For 15 successes and 8 failures, the normal approximation to the null distribution places the upper two-sided 5% significance cutoff near 0.75; reported p-values use permutations.

Table 1: **The checks at a glance.** Extra cost is per replanning step, in addition to the planner's own computation. Results are for the benchmark's default configuration (Section 5).

Check	Question	Reference	Extra cost per step	Result here
1	Does the plan move the imagined outcome more than the seed?	$R = 1$: equal effects	$k-1$ reseeded and 3 perturbed rollouts	$R = 0.81$ (VLM), 0.57 (heuristic)
2	Is the imagination closer to its own plan's true future?	positive margin	4 simulator renders	margin 1.5% of e_{self}; $p = 0.14$
3	Does the evaluator reproduce its decision?	high agreement	5 evaluator samples	mean $q = 0.24$; $q = 0$ at 43% of steps

4 THREE CHECKS

Check 1: plan versus seed. Let $s_0 = S(W(o, a, \xi_0))$ be the score of the executed plan under the reference seed. The seed effect and the plan effect at a step are

$$\Delta_{\text{seed}} = \frac{1}{k-1} \sum_{i=1}^{k-1} \big|S(W(o, a, \xi_i)) - s_0\big|, \qquad \Delta_{\text{plan}} = \frac{1}{3} \sum_{m=1}^{3} \big|S(W(o, a'_m, \xi_0)) - s_0\big|, \tag{2}$$

and the check reports $R = \mathbb{E}[\Delta_{\text{plan}}]/\mathbb{E}[\Delta_{\text{seed}}]$. Both effects use the same reference, and plan perturbations hold the seed fixed. The ratio measures sensitivity of the score to the action prompt relative to seed variation. A prompt change may also alter generation in ways unrelated to the physical action, so the numerator is an operational prompt effect. At $R < 1$, reseeding changes the score more on average than a one-primitive plan change. This provides a scale comparison for single-rollout planning, rather than a universal threshold for correct plan ranking.

An offline variant measures the displacement of the final frame in pixels for two whole plans (straight versus turning) on real frames, which allows configurations to be compared in minutes (Appendix C).

Check 2: own plan versus other plans. Let d be LPIPS between final frames, $o^\star(a)$ the true future under plan a, and $\hat{o}(a)$ the imagination, averaged over the five seeds. Define $e_{\text{self}} = d(\hat{o}(a), o^\star(a))$ and $e_{\text{cross}} = \frac{1}{3}\sum_m d(\hat{o}(a), o^\star(a'_m))$. A positive margin $e_{\text{cross}} - e_{\text{self}}$ indicates that the prediction is closer to its own plan's future under this metric. Its magnitude also depends on how different the simulator futures are.

Check 3: decision reproducibility. Let q be the fraction of the five re-queried samples that equal the executed plan. It estimates the probability that the evaluator assigns to its own decision. If the evaluator chose uniformly among the 3^H plans of the executed length, $\mathbb{E}[q] = 3^{-H}$. The benchmark samples the evaluator at temperature 1.0, and we keep that setting. Here q measures agreement under the deployed sampling policy. It is separate from the probability that the executed plan succeeds.

Confidence signals. We evaluate four training-free signals, each averaged over the steps of an episode and oriented so that larger values mean more confidence: **SC-k**, the negated standard deviation of the k seed scores; **APS**, the negated plan effect Δ_{plan}; **LL**, a likelihood proxy (the negated mean change of the latent over the last denoising steps); and **VLM-C**, the agreement q. The target is episode success.

5 RESULTS

5.1 CHECK 1: THE PLAN MOVES THE SCORE LESS THAN THE SEED

Under the VLM planner the mean seed effect is 0.025 and the mean plan effect 0.021, a ratio of 0.81 (95% CI $[0.69, 0.95]$, 326 steps). The plan effect is the smaller of the two at 64% of steps and in 20 of 23 episodes (sign-flip $p = 0.003$). Under the heuristic planner the ratio is 0.57 ($[0.54, 0.60]$,

Table 2: **Offline Check 1.** Mean absolute displacement of the final imagined frame (384^2 pixels, intensities in $[0, 1]$) when only the plan changes (straight versus turning left or right, common seed) and when only the seed changes (straight plan, two seeds), conditioned on two goal images from the benchmark. The shipped configuration is the first row.

Configuration	Plan effect	Seed effect	Ratio	Seconds per rollout
1280×704, padded prompt	0.155	0.171	0.90	60.6
1280×704, unpadded	0.131	0.164	0.80	53.3
640×352, padded	0.069	0.197	0.35	21.8
640×352, unpadded	0.031	0.245	0.13	22.4

210 steps) and the plan effect is smaller at 72% of steps and in all 10 episodes ($p = 0.002$). For the heuristic planner, this is the score used to rank candidates. The comparison quantifies seed sensitivity around the executed plan; it does not measure ranking accuracy over the full candidate set.

Table 2 repeats the comparison offline with whole plans, using the harness's own prompt construction on real HM3D frames. All four measured ratios are below one. The default configuration has the largest ratio, 0.90. Removing *forward* padding and halving resolution together reduce it to 0.13, while generation time falls by a factor of three.

5.2 Check 2: the imagination barely prefers its own plan's future

Over 326 steps, $e_{\text{self}} = 0.669$ and $e_{\text{cross}} = 0.679$. The margin of 0.010 LPIPS is 1.5% of the imagination's error. The margin is positive at 184 steps. Averaging within episodes gives a mean margin of 0.006, positive in 15 of 23 episodes (sign-flip $p = 0.14$). Descriptive step-level correlations with elapsed episode fraction are -0.11 for the margin and 0.01 for e_{self}.

For comparison with Hansen & Wang (2026), who call a model action-marginalized when the ratio of its error under shuffled actions to its error under true actions is at most 1.1, the analogous ratio here is $e_{\text{cross}}/e_{\text{self}} = 1.015$ (episode-clustered 95% CI $[1.003, 1.026]$). The two quantities are not identical: our alternative plans differ in one primitive rather than being drawn from other episodes, which pulls the ratio toward one, and we compare one imagination against two true futures instead of two imaginations against one. We report the ratio as a measure of relative fidelity, rather than applying their threshold to a different intervention.

5.3 Check 3: the evaluator rarely reproduces its decision

Across 324 re-queried steps, the executed plan appears in none of the five new samples at 43% of steps and in all five at 3% (Figure 1b); the mean agreement is $q = 0.24$. The evaluator is far from uniform, since uniform choice would give at least one match at 7% of steps against the 57% observed. The observed mean agreement corresponds to about one match in four independent re-queries of the evaluator at temperature 1.0. This concerns the evaluator with its original inputs held fixed, rather than a rerun of candidate generation or video sampling.

The reported variance decomposition places 92% of the variation in q within episodes. Descriptive rank correlations are 0.14 with plan length and 0.19 with the imagined score. Steps with no matching sample have similar fidelity to other steps (AUC = 0.54).

Greedy decoding would make the evaluator deterministic for fixed inputs. It would change the decision policy, however, and its task performance is not measured here. Agreement under temperature sampling should therefore be reported separately from the accuracy of either policy. Appendix E relates the agreement statistic to the evaluator's distribution over plans.

5.4 Confidence signals

Table 3 reports episode-level discrimination for the VLM planner. The four training-free signals have AUCs between 0.47 and 0.66, with permutation p-values between 0.22 and 0.93. VLM-C has the largest point estimate. SC-k and APS measure variation in rollout scores; the results do not establish

Table 3: **Episode-level discrimination** under the VLM planner (15 successes, 8 failures). Every signal is oriented so that larger values mean more confidence; $\mathrm{AUC} > 0.5$ means that successful episodes receive more confidence. The oracle uses the true future and is not available to a deployed agent. The approximate upper null cutoff is 0.75.

Signal	Definition (averaged over steps)	AUC	p
SC-k	$-$std of scores over $k = 5$ seeds	0.47	0.82
APS	$-\Delta_{\mathrm{plan}}$	0.59	0.50
LL	likelihood proxy from the denoiser	0.52	0.93
VLM-C	evaluator agreement q	0.66	0.22
Oracle	$-e_{\mathrm{self}}$ (all steps)	0.12	0.001
Oracle	$-e_{\mathrm{self}}$ (first 8 steps)	0.20	0.02
Oracle	$-e_{\mathrm{self}}$ (first 5 steps)	0.24	0.05

that this variation predicts episode success. The simulator-based fidelity oracle shows a stronger association in the opposite direction.

5.5 Fidelity to the truth is inversely related to success

The oracle assigns higher confidence to failed episodes than to successful ones ($\mathrm{AUC} = 0.12$; Table 3). Figure 1c relates this association to episode length. All eight failed episodes end at the 21-step budget, and their imagined and true futures agree more closely than those of successful episodes. This is what one expects when little in the view changes, as when an agent turns in place or stands before an obstacle, whereas an agent that moves through the scene produces changes that the model must predict. Across episodes, e_{self} falls with episode length (Spearman $\rho = -0.65$). Restricting every episode to its first five or eight steps removes the difference in length and weakens the inversion (AUC 0.24 and 0.20) without reversing it, which is consistent with a role for scene dynamics beyond episode duration. Scene identity cannot be separated from outcome here: none of the 19 scenes appears with both a success and a failure. Any gate based on the distance between imagination and reality should be tested with length and scene controlled.

5.6 The two planners

The VLM planner succeeds in 15 of 23 episodes and the heuristic planner in none of 10; every heuristic episode ends at the step budget. On the five episodes that both ran, the VLM planner succeeds in three and the heuristic planner in none. The heuristic planner uses the score in Eq. 1, whose seed sensitivity is measured by Check 1. The outcome comparison is consistent with the value of VLM-based planning reported by Zhang et al. (2026). Separating the contributions of the VLM and the world model requires a matched ablation; the two planner branches alone do not identify them.

6 Discussion

Reporting the checks. For each world model and configuration, report the plan-to-seed ratio with an interval over episodes. Where simulator branching is available, compare predictions with futures from both the executed plan and perturbed plans. Report evaluator agreement at the decoding settings used in deployment. Finally, test confidence signals against task outcomes, with episode length and scene included in the analysis. These measurements answer distinct questions and should be interpreted together.

Sampling and configuration. Averaging independent rollouts per candidate can reduce variation in an estimated mean score. The computational cost grows with the number of rollouts, and the resulting ranking should be evaluated directly. Prompt construction and image resolution also affect the comparison: the four offline configurations span plan-to-seed ratios from 0.13 to 0.90.

Interpreting confidence. A low plan-to-seed ratio identifies weak local score sensitivity relative to sampler variation. It does not by itself determine task performance or preclude calibrated predictions of success. Similarly, a high evaluator agreement can accompany either a correct or an incorrect decision. The episode-level analysis supplies the outcome comparison needed to interpret both quantities.

Scope of evaluation. We study one task, one world model, and two planners, with 33 episodes. The checks are defined for any generator with a seed, but whether other models pass them is an empirical question that our data do not answer; the offline check in Table 2 uses two conditioning frames per configuration. Check 1 uses one-primitive perturbations, the smallest change between candidate plans, and larger changes should produce larger plan effects. Distances are measured with LPIPS on final frames. We did not record the distance between the true futures of the executed and perturbed plans, so the margin of Check 2 is expressed relative to the imagination's error rather than to the size of the true plan effect. True futures were rendered for the VLM planner only. The estimator comparison estimates discrimination over the 23 VLM episodes.

Outlook. Trained, trajectory-conditioned navigation models (Bar et al., 2025) are the natural positive control for Check 1, and methods that learn calibrated uncertainty (Mei et al., 2025) can be evaluated against it. For deployment, a signal can be calibrated on held-out episodes to define an execute-or-defer policy. Conformal risk control (Angelopoulos et al., 2024) is one option when its exchangeability and loss assumptions are satisfied.

REPRODUCIBILITY STATEMENT

Appendix A specifies the statistical procedures, Appendix B describes the instrumentation, and Appendix C gives the offline protocol. The source package contains the manuscript, the figure, and the plotting script. The benchmark, world model, and VLM are publicly available.

LLM USAGE STATEMENT

GPT was used for writing polish.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, 2018.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, 2021.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Julia Berger, Bernd Frauenknecht, Sebastian Trimpe, and Bastian Leibe. Biased dreams: Limitations to epistemic uncertainty quantification in latent space models. *arXiv preprint arXiv:2604.25416*, 2026.

Jake Bruce, Michael Dennis, Ashley Edwards, et al. Genie: Generative interactive environments. In *International Conference on Machine Learning*, 2024.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, 2018.

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In *Advances in Neural Information Processing Systems*, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

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

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

Nicklas Hansen and Xiaolong Wang. Hallucination in world models is predictable and preventable. *arXiv preprint arXiv:2606.27326*, 2026.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI Conference on Artificial Intelligence*, 2018.

Ziqi Huang, Yinan He, Jiashuo Yu, et al. VBench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, 2019.

Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Pathdreamer: A world model for indoor navigation. In *IEEE/CVF International Conference on Computer Vision*, 2021.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Zhiting Mei, Tenny Yin, Micah Baker, Ola Shorinwa, and Anirudha Majumdar. World models that know when they don't know: Controllable video generation with calibrated uncertainty. *arXiv preprint arXiv:2512.05927*, 2025.

NVIDIA, Niket Agarwal, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. In *Conference on Robot Learning*, 2023.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied AI research. In *IEEE/CVF International Conference on Computer Vision*, 2019.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Yang Yu, Shiyuan Zhang, Yifei Sheng, Haoxiang Ren, and Haoxin Lin. How should world models be evaluated for embodied decision-making? A decision-making-centric position. *arXiv preprint arXiv:2606.15032*, 2026.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoğlu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M. Patel, Paul Pu Liang, Daniel Khashabi, Cheng Peng, Rama Chellappa, Tianmin Shu, Alan Yuille, Yilun Du, and Jieneng Chen. World-in-world: World models in a closed-loop world. In *International Conference on Learning Representations*, 2026.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

A Statistical procedures

Confidence intervals for ratios of means are percentile intervals from 10^4 bootstrap resamples of episodes; all steps of a resampled episode enter together. Per-episode tests are two-sided sign-flip permutation tests on per-episode means (2×10^4 random sign patterns). Steps are treated as dependent; step-level correlations are descriptive. AUC is the Mann–Whitney statistic, and rank correlations are Spearman's ρ; both have two-sided permutation p-values from 5×10^3 permutations. The approximate two-sided 5% significance cutoff for AUC is $0.5 + 1.96 s_0$, using the null standard deviation $\sqrt{(n_1 + n_0 + 1)/(12\, n_1 n_0)} = 0.13$ for $n_1 = 15$, $n_0 = 8$. The intra-class correlation of q is the between-episode variance of episode means divided by the total.

B Instrumentation

The hooks run inside the benchmark's solver immediately after the plan is chosen and before the first primitive is executed. Re-imagination calls the world model directly, so the imagined futures that the evaluator saw remain in the agent's state; all files written by the hooks go to a separate directory. The seeds are $\xi_i = \xi_0 + 1000\, i$ with ξ_0 a deterministic hash of the episode and step. True futures are rendered by saving the simulator state, executing each plan, recording the observations, and restoring the saved state. The evaluator re-query calls the benchmark's own decision function with the same state and arguments as the original query and requests five samples instead of one; its output is compared with the executed plan as a sequence of primitives.

C Offline protocol

The conditioning frames are the first two goal images of the benchmark's image-goal navigation set, and prompts are built with the harness's own function and padding. Three plans are compared: three *forward* steps, three left turns, and three right turns; after padding with *forward*, the turning prompts share 10 of their 13 action phrases with the straight prompt, and without padding they contain only the three actions. For each configuration and frame, the plan effect is the mean absolute difference between the final frame of the straight plan and those of the two turning plans at a common seed, and the seed effect is the difference between two seeds of the straight plan. Frames are resized to 384^2, the resolution the planner consumes. The plan effect is the mean absolute pixel difference between plans at matched seed, and the seed effect the difference between seeds at matched plan. Timing is wall-clock on one H100.

D EVALUATOR AGREEMENT BY PLAN LENGTH

Plan length H	Steps	Mean q	Steps with ≥ 1 match	Uniform: ≥ 1 match
1	5	0.04	20%	87%
2	16	0.15	56%	45%
3	17	0.32	65%	17%
4	32	0.11	38%	6%
5	254	0.27	59%	2%

E AGREEMENT AND THE DISTRIBUTION OVER PLANS

Let $p_t(a)$ be the evaluator's distribution over plans for fixed inputs at step t. Conditional on the executed plan A_t, the repeated-query statistic satisfies $\mathbb{E}[q_t \mid A_t] = p_t(A_t)$. If the original plan and the re-queries are independent samples from the same distribution, then

$$\mathbb{E}[q_t] = \sum_a p_t(a)^2. \tag{3}$$

Thus the mean agreement estimates a collision probability. Writing $p_{t,\max} = \max_a p_t(a)$ gives $p_{t,\max}^2 \leq \sum_a p_t(a)^2$ and, after averaging over steps,

$$\mathbb{E}_t[p_{t,\max}] \leq \sqrt{\mathbb{E}_t\left[\sum_a p_t(a)^2\right]}. \tag{4}$$

Substituting the observed mean agreement, 0.24, gives a plug-in value of about 0.49. This is a bound on concentration under the sampling policy, subject to the same-distribution assumption and estimation error. It is neither a probability of task success nor an evaluation of greedy decoding.

F COMPUTE

One Wan2.2-TI2V-5B rollout at 1280×704 with 50 denoising steps takes 64.8 s on an H100 and about 273 s on an L40S. An instrumented episode costs about four H100 GPU-hours, most of it in the extra rollouts of Check 1 and Check 2. The VLM runs on one H100 through vLLM with 4-bit AWQ weights.