FRAMETRUST: OUTCOME AGREEMENT FOR RELIABLE VIDEO IMAGINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As video models evolve into implicit world models, autonomous agents can "imagine before acting"—yet visual plausibility does not guarantee predictive reliability, risking irreversible errors. To bridge this gap, we introduce FrameTrust, a training-free visual self-consistency framework that isolates true outcome disagreement from mere path diversity by mapping diverse rollouts to discrete terminal states via task readers. For stochastic environments, a simulator-relative entropy contrast effectively disentangles genuine world randomness from model hallucinations. To mitigate extreme video sampling costs, an intermediate diffusion readout screens divergent futures before all generation steps are spent. These signals drive an intent gate that dynamically routes the agent toward execution, replanning, or a conservative fallback. Across spatial, physical, memory, and interface environments, FrameTrust significantly improves error ranking and reduces denoising work through early screening. Crucially, our gate reduces irreversible GUI execution errors by 59–64% relative to blind imagination without compromising task success, establishing a rigorous foundation for safe, imagination-driven autonomous action.

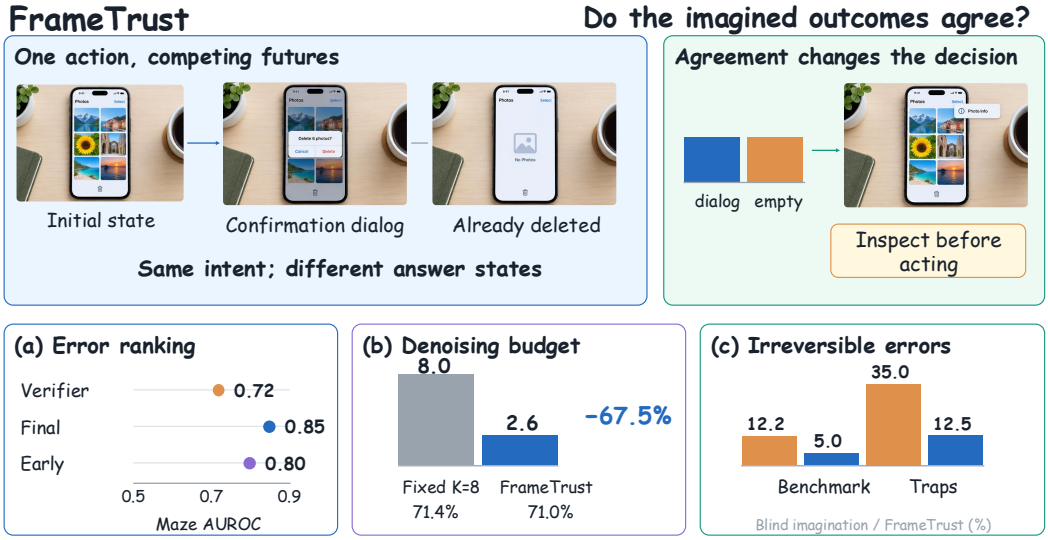

Figure 1: FrameTrust connects agreement to action. The lower panels summarize error ranking, denoising cost, and GUI decisions; the compute comparison retains the corresponding accuracy values.

1 INTRODUCTION

The deployment of generative video models is undergoing a profound paradigm shift: from tools for creative synthesis to cognitive engines for autonomous agents. By predicting the visual consequences of an action before it is executed, video generators now serve as implicit world models, enabling agents to plan across web interfaces, operating systems, and embodied environments (Gu et al., 2024; Du et al., 2023; Yang et al., 2024). However, this predictive capability harbors a critical vulnerability. A video model might imagine a perfectly plausible and visually stunning future—such as a confirmation dialog appearing after a "delete" command—that is factually incorrect for

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.

the specific environment. If an agent acts upon this hallucinated future, visual plausibility directly translates into catastrophic, irreversible errors. To safely deploy video-based reasoning, a fundamental question must be answered: *How can an agent quantify its epistemic trust in an imagined future before taking action?*

In language models, assessing reliability often relies on self-consistency—sampling multiple reasoning paths and aggregating their answers (Wei et al., 2022; Wang et al., 2023; Kuhn et al., 2023). Extending this principle to video generation, however, poses distinct and non-trivial challenges. First, video operates in a high-dimensional, continuous space where pixel variance does not equate to semantic divergence. Two clips may differ drastically in appearance or object trajectory while achieving the exact same task outcome (path vs. outcome divergence). Second, video environments exhibit genuine aleatoric uncertainty. In stochastic physics scenarios or dynamic GUIs, multiple valid futures can legitimately unfold from the same action. A naive measurement of outcome variation cannot distinguish whether the generator is confused (epistemic uncertainty) or merely capturing the natural stochasticity of the world. Finally, each imagination step requires a computationally expensive denoising rollout, rendering brute-force multi-sample consensus prohibitively costly for real-time decision-making.

To overcome these fundamental barriers, we introduce **FrameTrust**, a visual self-consistency framework that transforms diverse generated videos into actionable, reliable evidence for agentic planning. Rather than comparing raw trajectories or pixel distributions, FrameTrust shifts the consistency paradigm to the *outcome space*. We employ task readers to abstract complex video rollouts into categorical terminal states, marginalizing out irrelevant path variations. To disentangle genuine environmental randomness from model hallucination, we introduce a *simulator-relative entropy contrast*, which calibrates the generator's outcome disagreement against the expected aleatoric variation of the environment.

Furthermore, FrameTrust addresses the steep computational cost of video imagination through an intermediate soft-readout mechanism. By measuring the mutual information between sampling seeds and early-stage denoising estimates, our framework detects emerging outcome disagreements before a rollout is fully synthesized. This enables adaptive early screening, allowing the agent to abort highly uncertain imaginations at a fraction of the computational cost. Ultimately, a sequential mode-identification rule and an intent gate route the agent's decision flow: proceeding with execution when imagination is consistent, triggering replanning upon confident contradiction, or resorting to a conservative fallback (e.g., human assistance) when the imagined futures remain unresolved.

We comprehensively evaluate FrameTrust using frozen open-weights generators (Wan2.2 and CogVideoX) and a closed-API model (Veo 3.1) across diverse testbeds: procedural mazes, rigid-body physics, cup-swap memory, GUI transitions, and open-ended generation bridges. Our results demonstrate that outcome abstraction consistently outperforms standard baseline uncertainties in error ranking. More importantly, when deployed in an Android GUI agent, the FrameTrust intent gate reduces irreversible execution errors by 59–64% relative to blind imagination, without degrading overall task success. FrameTrust seamlessly bridges the gap between generative video reasoning and safe execution, establishing a principled foundation for when and how an agent should trust its own imagination.

Our key contributions are:

- A visual self-consistency framework that isolates task-critical outcome agreement from irrelevant spatial and trajectory variations using task readers.
- A simulator-relative entropy contrast to disentangle genuine environment stochasticity from model uncertainty, alongside a seed–answer mutual-information readout for early compute-saving screening during the diffusion process.
- A sequential sampling procedure and intent gate that operationalizes video agreement into actionable agent routing: selective execution, replanning, or conservative action.
- Extensive empirical validation showing superior error ranking and up to a 64% reduction in irreversible agent errors at favorable accuracy–compute trade-offs, requiring no additional training for the video backbone.

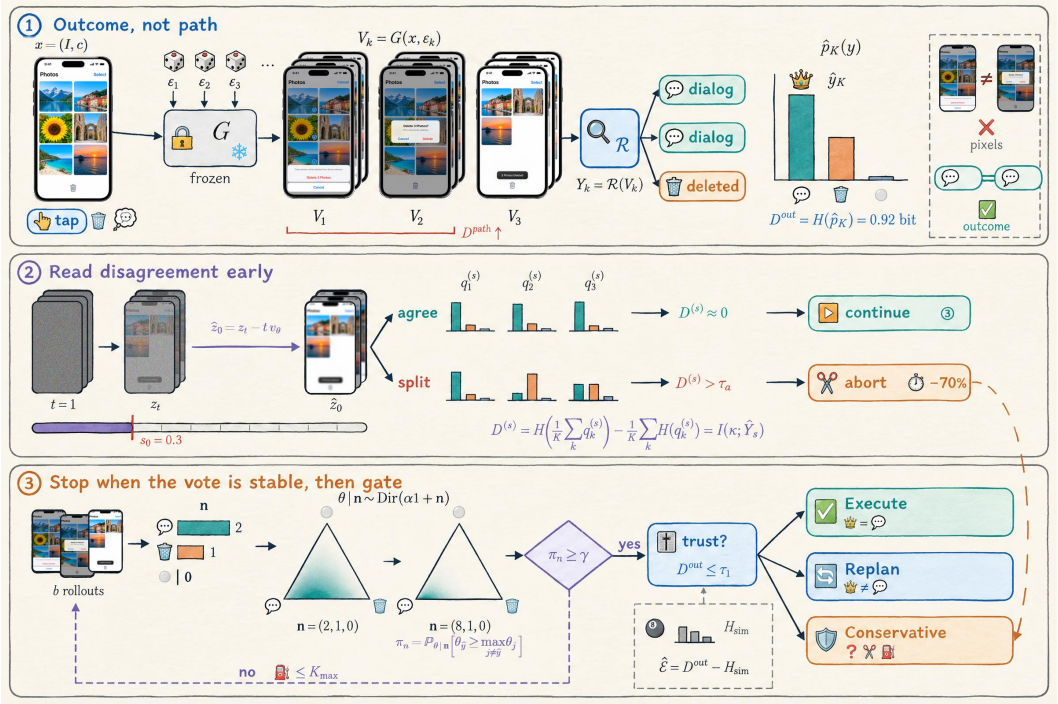

Figure 2: FrameTrust overview. A frozen generator samples rollouts that a reader maps to outcomes, not paths. Intermediate readouts screen disagreement early; a stable count posterior then gates execution, replanning, or deferral.

2 Related Work

Video reasoning and imagined environments. Zero-shot video reasoning and its benchmarks reveal capabilities in spatial search, physics, and memory (Wiedemer et al., 2025; Guo et al., 2026; Tong et al., 2025); denoising analysis motivates latent-trajectory ensembling (Wang et al., 2026). Video policies, simulators, and world models support planning across embodied, web, and interface tasks (Du et al., 2023; Yang et al., 2024; Du et al., 2024; Bar et al., 2025; Ha & Schmidhuber, 2018; Hafner et al., 2023; Gu et al., 2024; Rivard et al., 2025; Qin et al., 2025). Video-T1 and VISTA use feedback for selection and refinement (Liu et al., 2025; Long et al., 2025). FrameTrust supplies an outcome-agreement signal and a decision procedure that can complement improved generators and external judges.

Disagreement, calibration, and adaptive sampling. Predictor disagreement underlies deep ensembles, dropout-based uncertainty, and Bayesian active learning (Lakshminarayanan et al., 2017; Gal & Ghahramani, 2016; Houlsby et al., 2011; Kendall & Gal, 2017; Depeweg et al., 2018). Language methods use answer voting, semantic equivalence, sampled consistency, and adaptive stopping (Wang et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024; Manakul et al., 2023; Aggarwal et al., 2023; Li et al., 2024; Snell et al., 2024; Brown et al., 2024). Verbal confidence, structured-prediction uncertainty, and pixel-wise diffusion uncertainty provide related signals (Kadavath et al., 2022; Xiong et al., 2024; Malinin & Gales, 2021; Kou et al., 2024). We study agreement across seeds of one frozen generator, with task readers and partial-generation access, using selective prediction and held-out calibration to assess its decision value (Geifman & El-Yaniv, 2017; El-Yaniv & Wiener, 2010; Hendrycks & Gimpel, 2017).

3 Outcome Agreement for Video Imagination

3.1 A generator-induced answer law

An instance $x = (I, c)$ contains an initial observation and instruction. A frozen generator maps a seed ϵ_k to video $V_k = G(x, \epsilon_k)$, and a reader maps that video to an answer $Y_k = \mathcal{R}(V_k) \in \mathcal{Y}$. With

$M = |\mathcal{Y}|$, the population law and its empirical estimate are

$$p_G(y \mid x) = \mathbb{P}_\epsilon[\mathcal{R}(G(x, \epsilon)) = y], \qquad \widehat{p}_K(y \mid x) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\{Y_k = y\}. \tag{1}$$

The prediction is the plurality $\widehat{y}_K = \arg\max_y \widehat{p}_K(y)$ and the outcome disagreement is $D^{\text{out}} = H(\widehat{p}_K) = -\sum_y \widehat{p}_K(y) \log_2 \widehat{p}_K(y)$. The benchmark compares this prediction with the environment's modal outcome $y^\star$. Outcome entropy is an empirical reliability signal; a separate calibration split maps it to error risk when a probability is needed. Appendix A derives conditions linking answer concentration to stable plurality identification.

3.2 Readers separate outcome from path variation

Programmatic readers track colors and geometry on procedural tasks, returning an exit, bin, stability state, or revealed cup. Structured readers classify terminal interface states against a fixed schema with a frozen vision-language model. Open-ended readers extract canonical prompt-constrained objects, counts, attributes, and relations. Invalid outcomes and unresolved readouts remain explicit in the evaluation record. Reader error is audited against independent labels because applying the same reader to generated and real observations does not guarantee that its mistakes cancel.

We distinguish D^{out} from path diversity D^{path}, measured through maze-state edit distances, dynamic time warping of object tracks, or intermediate-frame feature distances (Sakoe & Chiba, 1978). Different valid paths can imply one answer. The reader provides the abstraction needed to marginalize this nuisance variation, as semantic equivalence does for language (Kuhn et al., 2023). Appendix C gives the extraction contract, an entropy perturbation bound for reader noise, and a sufficient condition for preserving the answer mode.

3.3 Agreement and sampling stability

The link between agreement and a reliable vote can be made precise. Let y_G be the unique mode of the generator's readout law, and let $\Delta = \min_{j \neq y_G}[p_G(y_G) - p_G(j)]$. For independent seeds, pairwise count differences lie in $[-1, 1]$, so a concentration bound gives

$$\mathbb{P}(\widehat{y}_K \neq y_G) \leq (M - 1) \exp(-K\Delta^2/2). \tag{2}$$

The bound describes identification of the model's mode. Across task instances, environment error is consequently bounded by model-mode error plus sampling instability:

$$\mathbb{P}(\widehat{y}_K \neq y^\star) \leq \mathbb{P}(y_G \neq y^\star) + \mathbb{E}_x\left[\min\{1, (M-1)e^{-K\Delta(x)^2/2}\}\right]. \tag{3}$$

This separates what additional imagination can stabilize from what a decision gate must assess. Increasing K reduces the sampling term when answers are separated; calibrated abstention targets instances on which a stable model answer is still unreliable. Appendix A gives the derivation and the relationship between population entropy and the modal gap.

The reader also changes the available margin. If its probability of changing the ideal video answer is ϵ_R, coupling bounds the induced distribution shift by ϵ_R in total variation. An ideal modal gap $\Delta_Z > 2\epsilon_R$ is sufficient to preserve the answer mode after reading. This gives a concrete reason to audit task-relevant readouts rather than rely only on general image-recognition quality: errors near a close decision can matter more than the same number of errors on well-separated outcomes. The reader-noise study tests this component independently of the generator, while the sequential rule uses observed counts instead of requiring the unknown population gap.

3.4 A task-relative entropy contrast

For stochastic tasks, let $p^\star_\nu$ be the outcome law under a specified distribution of environment perturbations. Cross entropy obeys

$$\mathbb{E}_{Y \sim p^\star_\nu}[-\log_2 p_G(Y \mid x)] = H(p^\star_\nu) + \mathrm{KL}_2(p^\star_\nu \| p_G). \tag{4}$$

Table 1: Testbeds. Instances are evaluation counts (adaptation counts in parentheses); $|\mathcal{Y}|$ includes an *invalid* outcome where applicable. The aleatoric baseline H_{sim} is available where a simulator generates the task.

Testbed	Skill probed	Instances	$\lvert\mathcal{Y}\rvert$	Reader	Ground truth	H_{sim}
Maze	spatial search, constraint	2,400 (20k)	4	programmatic	unique connected exit	$= 0$
Physics	dynamics, causality	1,800 (20k)	2 to 5	programmatic	simulator; 720 stochastic	✓
Cups	working memory	1,000 (10k)	3	programmatic	swap bookkeeping	$= 0$
GUI	interface dynamics	800 (30k)	5	structured VLM	real execution	$= 0$
T2V bridge	open-ended adherence	300	tuples	open-ended VLM	none (verifier, humans)	n/a

The KL term measures outcome-law mismatch. The operational score uses only the scalar entropy baseline,

$$\widehat{\mathcal{E}}(x) = D^{\text{out}}(x) - H_{\text{sim}}(x), \tag{5}$$

where H_{sim} is estimated from 32 simulator outcomes with the registered one-pixel position and one-percent velocity perturbations. Positive excess means one imagined outcome is not yet a basis for action; negative excess means agreement has fallen below the simulator's own variation and is not evidence of safety. On deterministic tasks the baseline is zero. A swapped mode preserves entropy while reversing the action, so equal entropy does not certify the outcome law. Appendix A supplies the identifiability analysis, and Appendix B fixes the perturbation protocol.

3.5 Reading disagreement before completion

For a flow generator with $z_t = (1-t)z_0 + t\epsilon$, the clean-latent estimate is $\widehat{z}_0 = z_t - tv_\theta(z_t, t, c)$; at the optimal regression velocity it equals $\mathbb{E}[z_0 \mid z_t, c]$ (Lipman et al., 2023; Liu et al., 2023). The corresponding noise-prediction diffusion estimate is $(z_t - \sigma_t\epsilon_\theta)/\alpha_t$ (Ho et al., 2020; Efron, 2011). At sampling progress s, we decode the estimate and obtain soft reader distributions $q_k^{(s)}$. Their between-seed disagreement is

$$D^{(s)} = H\left(\frac{1}{K}\sum_k q_k^{(s)}\right) - \frac{1}{K}\sum_k H(q_k^{(s)}) = I(\kappa; \widehat{Y}_s), \tag{6}$$

where κ selects an empirical seed uniformly and $\widehat{Y}_s \mid \kappa = k \sim q_k^{(s)}$. This generalized Jensen–Shannon identity (Lin, 1991) removes uncertainty shared by identical soft readouts and responds when seeds support different answers. It lies between zero and $\min(\log_2 K, \log_2 M)$. Final hard readouts recover D^{out}; intermediate disagreement need not grow monotonically.

The default screening position is $s_0 = .3$ of executed sampling steps, motivated by the supplied progress study and denoising-stage analysis (Wang et al., 2026). Progress is converted through the actual scheduler rather than assumed equal to a noise coordinate. Decoding and reading costs are included in wall-time accounting. Closed interfaces that expose only finished videos use the final-output variant.

3.6 Sequential sampling and the intent gate

Completed answers update a categorical count posterior $\theta \mid \boldsymbol{n} \sim \text{Dir}(\alpha\mathbf{1} + \boldsymbol{n})$. For current plurality $\widehat{y}$, the mode-identification statistic is

$$\pi_n = \mathbb{P}_{\theta|\boldsymbol{n}}\left[\theta_{\widehat{y}} \geq \max_{j \neq \widehat{y}} \theta_j\right]. \tag{7}$$

We draw batches until $\pi_n \geq \gamma$ or the sample cap is reached, following adaptive-consistency methods (Aggarwal et al., 2023; Li et al., 2024). The batch size is clipped to the remaining budget. The first batch can abort at s_0 when $D^{(s_0)} > \tau_a$; otherwise its existing sampler states continue to completion. Appendix G derives the posterior computation, controls its Monte Carlo error, and distinguishes mode identification from environment correctness.

After a stable plurality is identified, a calibrated threshold on final disagreement, or the registered simulator-relative score, determines whether to trust it. A trusted outcome matching intent supports execution; a trusted contradiction supports replanning; unresolved disagreement selects a conservative branch. Cap and early-screen abstentions also enter that branch. Thresholds are chosen on calibration data collected under the same sampling policy. On interfaces the branch uses a declared reversible probe or user assistance, with its cost and task outcome recorded.

Algorithm 1 FrameTrust: sample, read, decide

Require: instance and intent, frozen generator, reader, $b, K_{\max}, s_0, \tau_a, \alpha, \gamma, \tau_1$

1: Initialize completed-answer counts to zero
2: Screen the first batch at s_0 using Eq. (6); abstain if its score exceeds τ_a
3: Complete surviving samples and update answer counts
4: Draw additional batches, clipped to the cap, until Eq. (7) passes or the budget is exhausted
5: If mode confidence does not pass, return the conservative branch
6: Compute final disagreement and the optional simulator-relative contrast
7: Return execution, replanning, or conservative action according to trust and intent

Table 2: AUROC for predicting errors of the K-rollout plurality answer ($K = 8$; Veo 3.1 with $K = 4$). All signals use the same rollouts. Err. is the error rate of the plurality answer. Stochastic physics are the 720 scenes with $H_{\text{sim}} > 0.1$ bits. Macro rows average over the four testbeds (three for Veo). Uncertainty and API-access conventions are in Appendix E.

			Comparison signals					FrameTrust (ours)		
Generator	**Testbed**	**Err.%**	PixVar	CFG-sens	Resid	Verifier	Self-rep	D^{out}	$\mathcal{E}$	$D^{(0.3)}$
Wan2.2 TI2V-5B	Maze	28.6	0.57	0.63	0.55	0.72	0.56	**0.85**	**0.85**	0.80
	Physics, det.	18.8	0.54	0.61	0.53	0.60	0.55	**0.86**	**0.86**	0.81
	Physics, stoch.	31.5	0.55	0.58	0.52	0.56	0.53	0.70	**0.84**	0.78
	Cups	25.2	0.56	0.62	0.54	0.65	0.55	**0.83**	**0.83**	0.79
	GUI	20.7	0.58	0.64	0.56	0.69	0.57	**0.81**	**0.81**	0.77
Wan2.2	macro	24.9	0.56	0.62	0.54	0.65	0.55	0.82	**0.84**	0.79
CogVideoX1.5-5B	macro	27.8	0.55	0.61	0.54	0.64	0.55	0.81	**0.83**	0.78
Veo 3.1 (zero-shot)	macro	34.0	0.55	n/a	n/a	0.63	0.56	0.79	**0.81**	n/a

4 EXPERIMENTS

Setup. Table 1 lists the five settings. Maze answers come from the rendered connectivity graph; physics outcomes come from a rigid-body simulator; cup answers follow the registered swaps; and interface outcomes come from AndroidWorld execution (Bakhtin et al., 2019; Rawles et al., 2025). The open-ended bridge uses T2V-CompBench prompts (Sun et al., 2025). Appendix B gives the instance contracts and Appendix C the readers. The primary generator is Wan2.2-TI2V-5B; CogVideoX1.5-5B-I2V provides a second architecture (Wan et al., 2025; Wan Team, 2025; Yang et al., 2025). Rank-64 task LoRAs are adapted on disjoint data and then frozen (Hu et al., 2022). Open-model videos use 480×832 resolution, 49 frames, and 40 steps with $K = 8$ unless stated. Veo 3.1 uses its supported API configuration, zero-shot on 100 instances per procedural family with $K = 4$ (Google DeepMind, 2025); intermediate-state diagnostics are unavailable. Appendix D records the access and configuration distinctions.

Confidence signals and evaluation. All signals predict the error of the same plurality answer. Comparisons include pixel variance, guidance sensitivity, a re-noising residual, a fixed Qwen2.5-VL judge, and verbal confidence (Liu et al., 2025; Bai et al., 2025; Xiong et al., 2024). Appendix E defines their inputs, extra calls, calibration, and instance-level uncertainty. AUROC measures error ranking; selective prediction additionally needs the original score/error pairs. Compute is reported in executed denoising equivalents, with reader, decoder, and fallback costs accounted for separately in wall time.

Outcome abstraction improves error ranking. Table 2 and Figure 3 connect the reader to the central claim: disagreement after task abstraction ranks errors more effectively than the comparison signals. Abstraction lifts deterministic physics from .61 to .86 because different trajectories share one terminal state; on stochastic physics, outcome entropy reaches .70, and excess reaches .84 only after the simulator baseline is removed. The closed-model result supports final-output agreement at a different access level, with .79 outcome and .81 excess macro AUROC.

Path diversity and task randomness. The distinction between route and outcome prevents harmless motion variation from becoming an automatic rejection criterion. The physics-subset comparison in Table 3 and Figure 4 instead tests variation in the outcome itself: excess disagreement adds .19 AUROC on pegboards and .11 on multi-ball scenes, where the simulator baseline is largest. Negative

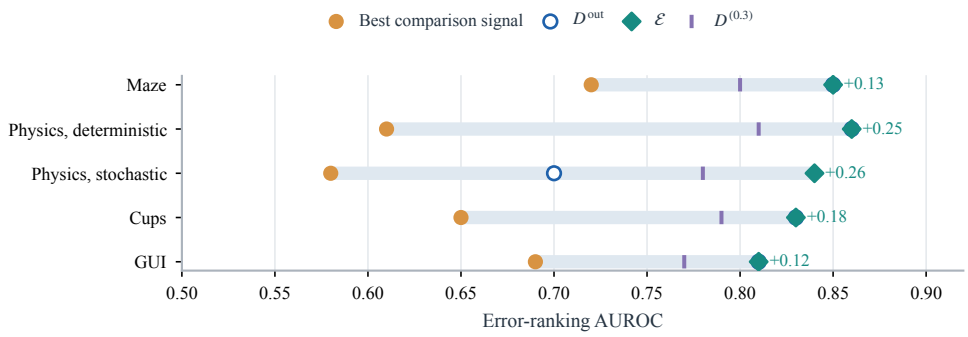

Figure 3: Error-ranking gains on Wan2.2. Paired AUROCs compare the strongest listed comparison signal with final, simulator-relative, and early agreement. Numbers at right give the excess-score gain.

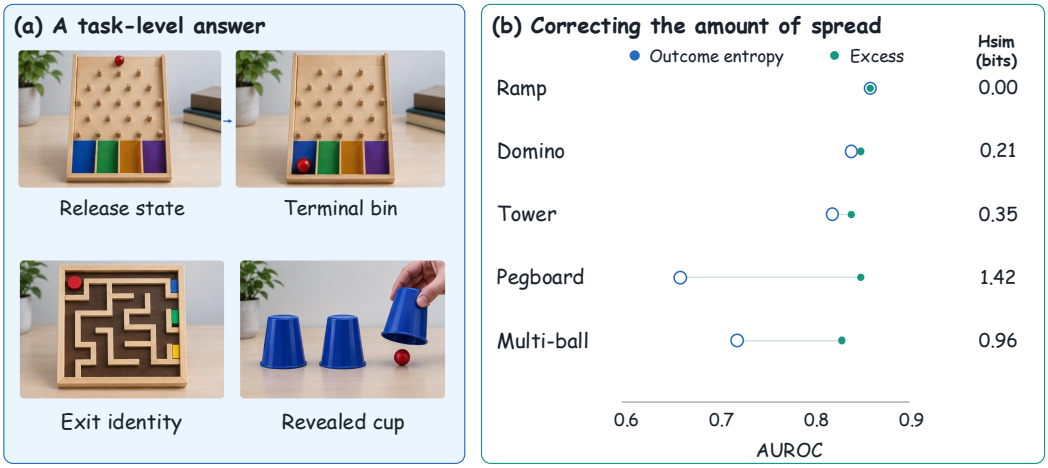

Figure 4: Task-level outcomes and simulator-relative disagreement. The right panel pairs physics-subset AUROCs with simulator entropies to show the effect of accounting for task randomness.

excess is retained as a separate underdispersion diagnostic. Appendix A explains why this signed contrast measures distribution width without identifying all distributional errors.

Trajectory ensembling and answer aggregation operate at different levels (Wang et al., 2026). Table 4 compares their composition at eight trajectory-equivalents, apart from the single-rollout reference. Four two-trajectory ensembles provide four votes, retaining an agreement signal while improving the video associated with each vote.

Early readout and sampling cost. Figure 5 tests the intermediate signal and its operational consequence. At 30% progress, the listed AUROC is within .05 of its final value across the four open-model settings. The screening frontier reaches 71.0% maze accuracy at 2.6 rollout-equivalents, compared with 71.4% at eight fixed rollouts: 67.5% less denoising work with a .4-point accuracy difference. At 4.4 equivalents it reaches 71.5%. These operating points include the declared fallback

Table 3: Physics subsets. AUROC and false-alarm rate at 90% error recall. H_{sim} is in bits.

Subset	$\bar{H}_{\text{sim}}$	AUROC D^{out}	AUROC $\mathcal{E}$	FA@90 D^{out}	FA@90 $\mathcal{E}$
Single ramp	.00	.86	.86	.31	.31
Domino	.21	.84	.85	.34	.32
Tower	.35	.82	.84	.37	.33
Pegboard	1.42	.66	**.85**	.61	**.24**
Multi-ball	.96	.72	**.83**	.52	**.29**

Table 4: Maze accuracy (%). The single-rollout reference uses one trajectory; other rows use eight. Acc.@80 uses 80% coverage.

Method	Accuracy (%) Acc.	Acc.@80
Single rollout	64.1	n/a
Traj-Ens (8)	68.4	n/a
Plurality ($K = 8$)	71.4	n/a
FrameTrust gate	71.4	79.1
FrameTrust + Ens	**72.0**	**79.8**

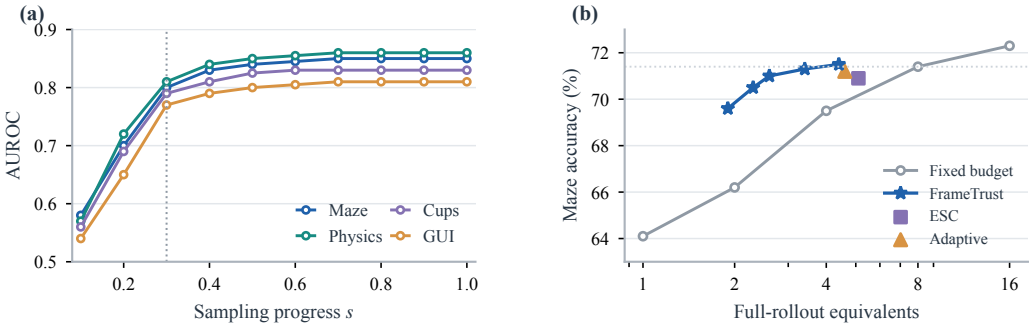

Figure 5: Intermediate agreement and compute allocation. (a) Error-ranking AUROC at registered sampling progress; Physics is the deterministic subset. (b) Maze accuracy against executed denoising work. ESC and Adaptive are early-stopping and adaptive-consistency baselines (Li et al., 2024; Aggarwal et al., 2023).

Table 5: GUI agent on AndroidWorld tasks with irreversible actions (benchmark) and constructed traps. Irrev. is the rate of irreversible actions whose real outcome contradicts the plan. Calls are generator rollout equivalents per gated decision; time is wall-clock per decision on one H100 with the reader included. Means over three runs, standard deviation ≤ 1.8 points.

	Benchmark (60)		Traps (40)			
Policy	Success↑	Irrev.↓	Success↑	Irrev.↓	Calls	Time (s)
No imagination	63.3	16.7	40.0	50.0	0	0
Blind imagination (K=1)	66.7	12.2	47.5	35.0	1.0	9
Plurality vote (K=8, no gate)	**68.9**	10.6	50.0	30.0	8.0	52
Verifier-gated (K=1)	66.1	10.0	50.0	27.5	1.0	14
FrameTrust gate (Alg. 1)	68.3	5.0	55.8	12.5	3.1	26
FrameTrust gate + logistic combiner	**68.9**	**4.4**	**56.7**	**11.7**	3.1	26

on aborted instances. Appendix G gives the complete stopping procedure and distinguishes posterior mode confidence from environment accuracy.

From confidence to GUI decisions. Table 5 evaluates the Qwen2.5-VL planner on benchmark tasks and constructed traps. The gate reduces irreversible errors by 59.0% and 64.3% relative to blind imagination while success rates increase by 1.6 and 8.3 points. Figure 6 explains how the policies can differ even when a generated screen looks plausible. Conservative probes are application-specific reversible operations; their costs and any human assistance belong in the decision trace (Appendix H).

Routing open-ended generation. Table 6 evaluates prompt-constrained tuple disagreement. Without a unique environment answer, the signal measures instability of constrained content and predicts judge rejection. Routing the top 30% of prompts recovers 81.25% of the adherence improvement of the full-verifier policy using no verifier calls; the hybrid recovers 93.75% while reducing verifier calls by 70%. Reader and rewrite costs remain part of both pipelines. Appendix I gives the tuple schema and the arithmetic behind these derived comparisons.

Joint decision quality. The GUI comparison also separates the cost of one imagined future from the cost of a fixed voting budget. Against eight-rollout voting, the gate uses 3.1 rollout-equivalents and 26 seconds per decision instead of eight and 52 seconds, while lowering irreversible errors on both splits. Benchmark success moves by only .6 points while trap success rises by 5.8, because the gate spends its refusals where an imagined screen and the environment disagree. The resulting operating point combines a lower generation budget with a decision rule that can decline or revise an unsafe plan. Its value is therefore assessed through complete task outcomes, including conservative actions, rather than only through accuracy on the subset accepted for execution.

Seed budget and reader robustness. Figure 7 tests the two measurement components underlying the signal. Additional seeds improve ranking with diminishing returns, while controlled reader corruption reduces it. Ranking flattens once the plurality is stable: maze AUROC moves only from

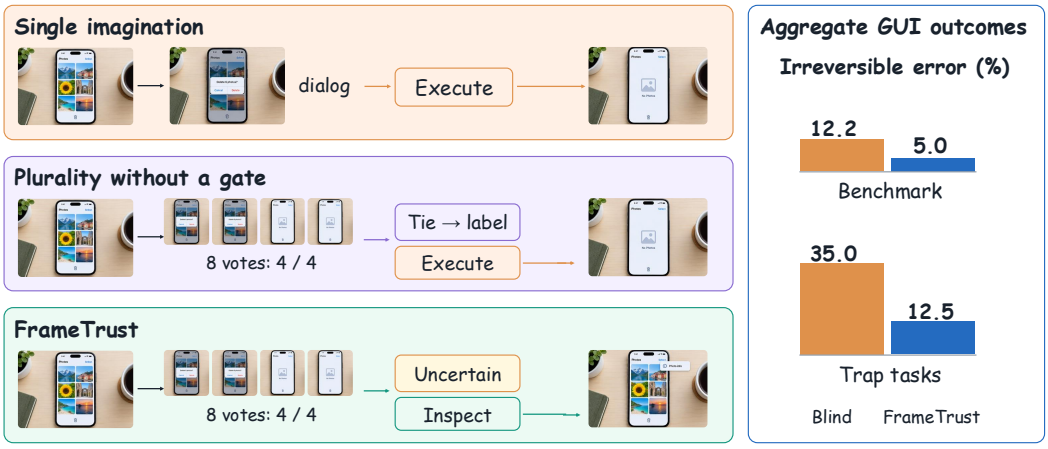

Figure 6: Agreement changes control flow. One imagined outcome, a tie-broken plurality, and an uncertainty-aware branch select different actions. The bars are the irreversible-error rates in Table 5.

Table 6: Bridge to open-ended generation (T2V-CompBench prompts, Wan2.2, $K = 4$). Left: how well each signal predicts verifier rejection and human preference variance. Right: a regeneration loop that rewrites and regenerates flagged prompts; adherence is the mean judge score of the final clips, computed post hoc for evaluation only.

Signal	AUROC↑	Spearman↑	**Routing policy**	Adherence↑	Verifier calls↓
Pixel variance	0.55	0.08	Single pass, no routing	0.664	0
CLIP feature variance	0.61	0.17	Random 30% rewrite	0.681	0
Verbalised confidence	0.58	0.12	Verifier filter + rewrite (all)	**0.712**	1,200
Tuple disagreement (FrameTrust)	**0.74**	**0.41**	FrameTrust top-30% rewrite	0.703	**0**
			FrameTrust top-30% + verifier on those	0.709	360

.85 at eight seeds to .86 at sixteen, so later samples mostly repeat a decision the earlier votes already support. The recorded cross-model correlation and top-20% overlap motivate transferred screening, whose deployment benefit remains a separate experiment. Appendix F supplies numerical summaries and controls that distinguish abstraction quality, model-mode stability, and decision utility.

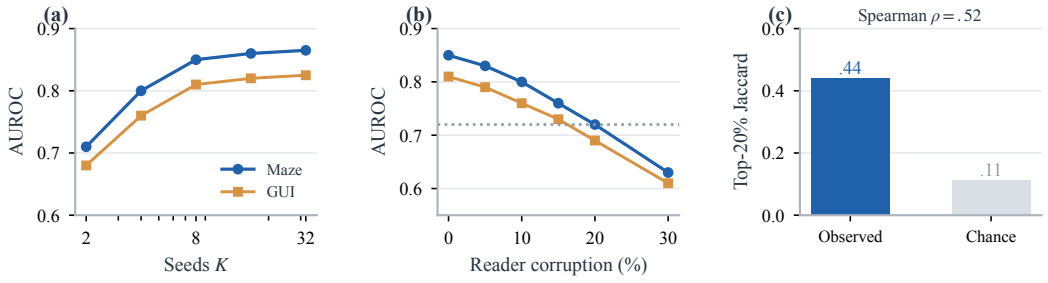

Figure 7: Measurement sensitivity. (a) Seed-budget sweep. (b) Controlled reader corruption, with the maze verifier reference dotted. (c) Cross-model overlap and rank-correlation summaries, compared with the random-subset overlap baseline.

5 Conclusion

We introduce FrameTrust, which uses outcome agreement to connect video imagination with agent decisions. Task readers, simulator-relative entropy contrasts, and intermediate readouts support adaptive sampling and selective execution with a frozen generator. The gains in error ranking, denoising efficiency, and GUI decision quality show how structured agreement can make generated futures more useful as evidence for action, while keeping sampling effort and the consequences of each decision explicit.

AI USE STATEMENT

In this work, authorized artificial intelligence tools were utilized exclusively for academic auxiliary purposes, including two primary scenarios as declared. First, AI writing polishing tools were adopted to refine the grammatical accuracy, logical fluency, academic rigor, and expression uniformity of the full manuscript, covering the introduction, related work, methodology, experiment analysis, and conclusion sections, without modifying the core research ideas, algorithm designs, experimental frameworks, empirical results, or original academic contributions of this paper. Second, AI-based retrieval and discovery systems were employed to assist in literature screening, keyword-based paper searching, and sorting of recent high-relevant studies in the field, which helped the authors comprehensively collect latest progress, sort out research gaps, and optimize the construction of the related work section. All AI-assisted processes are transparent and fully disclosed; all research designs, experimental implementations, result analyses, and academic arguments are independently completed and verified by the authors, and no AI tool was involved in data generation, model training, result fabrication, or academic reasoning. Detailed descriptions of AI assistance are supplemented in the appendix of this paper to ensure complete openness and compliance with ICLR 2027 submission regulations.

REFERENCES

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. Let's Sample Step by Step: Adaptive-Consistency for Efficient Reasoning and Coding with LLMs. In *Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://arxiv.org/abs/2305.11860.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. PHYRE: A New Benchmark for Physical Reasoning. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1908.05656.

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. URL https://arxiv.org/abs/2412.03572.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large Language Monkeys: Scaling Inference Compute with Repeated Sampling. *arXiv preprint arXiv:2407.21787*, 2024. URL https://arxiv.org/abs/2407.21787.

Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning. In *International Conference on Machine Learning*, 2018. URL https://arxiv.org/abs/1710.07283.

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. URL https://arxiv.org/abs/2302.00111.

Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B. Tenenbaum, Leslie Kaelbling, Andy Zeng, and Jonathan Tompson. Video Language Planning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.10625.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *International Conference on Machine Learning*, 2016. URL https://arxiv.org/abs/1506.02142.

Yonatan Geifman and Ran El-Yaniv. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1705.08500.

Google DeepMind. Veo 3.1. https://deepmind.google/models/veo/, 2025.

Yu Gu, Kai Zhang, Yuting Ning, Boyuan Zheng, Boyu Gou, Tianci Xue, Cheng Chang, Sanjari Srivastava, Yanan Xie, Peng Qi, Huan Sun, and Yu Su. Is Your LLM Secretly a World Model of the Internet? Model-Based Planning for Web Agents. *arXiv preprint arXiv:2411.06559*, 2024. URL https://arxiv.org/abs/2411.06559.

Ziyu Guo, Xinyan Chen, Renrui Zhang, Ruichuan An, Yu Qi, Dongzhi Jiang, Xiangtai Li, Manyuan Zhang, Hongsheng Li, and Pheng-Ann Heng. Are Video Models Ready as Zero-Shot Reasoners? An Empirical Study with the MME-CoF Benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9175–9184, 2026. URL https://arxiv.org/abs/2510.26802.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, 2018. URL https://arxiv.org/abs/1809.01999.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering Diverse Domains through World Models. *arXiv preprint arXiv:2301.04104*, 2023. URL https://arxiv.org/abs/2301.04104.

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1610.02136.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.11239.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian Active Learning for Classification and Preference Learning. *arXiv preprint arXiv:1112.5745*, 2011. URL https://arxiv.org/abs/1112.5745.

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.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1703.04977.

Siqi Kou, Lei Gan, Dequan Wang, Chongxuan Li, and Zhijie Deng. BayesDiff: Estimating Pixel-wise Uncertainty in Diffusion via Bayesian Inference. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.11142.

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. URL https://arxiv.org/abs/2302.09664.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1612.01474.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. Escape Sky-high Cost: Early-stopping Self-Consistency for Multi-step Reasoning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2401.10480.

Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Fangfu Liu, Hanyang Wang, Yimo Cai, Kaiyan Zhang, Xiaohang Zhan, and Yueqi Duan. Video-T1: Test-Time Scaling for Video Generation. *arXiv preprint arXiv:2503.18942*, 2025. URL https://arxiv.org/abs/2503.18942.

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.

Do Xuan Long, Xingchen Wan, Hootan Nakhost, Chen-Yu Lee, Tomas Pfister, and Sercan Ö. Arık. VISTA: A Test-Time Self-Improving Video Generation Agent. *arXiv preprint arXiv:2510.15831*, 2025. URL https://arxiv.org/abs/2510.15831.

Andrey Malinin and Mark Gales. Uncertainty Estimation in Autoregressive Structured Prediction. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2002.07650.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://arxiv.org/abs/2303.08896.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, Wanjun Zhong, Kuanye Li, Jiale Yang, Yu Miao, Woyu Lin, Longxiang Liu, Xu Jiang, Qianli Ma, Jingyu Li, Xiaojun Xiao, Kai Cai, Chuang Li, Yaowei Zheng, Chaolin Jin, Chen Li, Xiao Zhou, Minchao Wang, Haoli Chen, Zhaojian Li, Haihua Yang, Haifeng Liu, Feng Lin, Tao Peng, Xin Liu, and Guang Shi. UI-TARS: Pioneering Automated GUI Interaction with Native Agents. *arXiv preprint arXiv:2501.12326*, 2025. URL https://arxiv.org/abs/2501.12326.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Toyama, Robert Berry, Divya Tyamagundlu, Timothy Lillicrap, and Oriana Riva. AndroidWorld: A Dynamic Benchmarking Environment for Autonomous Agents. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2405.14573.

Luke Rivard, Sun Sun, Hongyu Guo, Wenhu Chen, and Yuntian Deng. NeuralOS: Towards Simulating Operating Systems via Neural Generative Models. *arXiv preprint arXiv:2507.08800*, 2025. URL https://arxiv.org/abs/2507.08800.

Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978. doi: 10.1109/TASSP.1978.1163055.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. *arXiv preprint arXiv:2408.03314*, 2024. URL https://arxiv.org/abs/2408.03314.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A Comprehensive Benchmark for Compositional Text-to-video Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2407.14505.

Jingqi Tong, Yurong Mou, Hangcheng Li, Mingzhe Li, Yongzhuo Yang, Ming Zhang, Qiguang Chen, Tianyi Liang, Xiaomeng Hu, Yining Zheng, Xinchi Chen, Jun Zhao, Xuanjing Huang, and Xipeng Qiu. Thinking with Video: Video Generation as a Promising Multimodal Reasoning Paradigm. *arXiv preprint arXiv:2511.04570*, 2025. URL https://arxiv.org/abs/2511.04570.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Wan Team. Wan2.2: Open and advanced large-scale video generative models. https://github.com/Wan-Video/Wan2.2, 2025.

Ruisi Wang, Zhongang Cai, Fanyi Pu, Junxiang Xu, Wanqi Yin, Maijunxian Wang, Ran Ji, Chenyang Gu, Bo Li, Ziqi Huang, Hokin Deng, Dahua Lin, Ziwei Liu, and Lei Yang. Demystifying Video Reasoning. *arXiv preprint arXiv:2603.16870*, 2026. URL https://arxiv.org/abs/2603.16870.

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. URL https://arxiv.org/abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2201.11903.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025. URL https://arxiv.org/abs/2509.20328.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.13063.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning Interactive Real-World Simulators. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.06114.

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.