PLANNING WITH GENERATED WORLDS: CALIBRATION-AWARE CROSS-WORLD VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative world models increasingly expose 3D environments and action-conditioned rollouts to downstream agents for planning. Yet visual realism, geometric fidelity, and rollout consistency do not establish that generated worlds retain task-relevant structure for reliable action. We formulate this gap as *generated-world actionability*: whether generated worlds support reliable decisions, when their hypotheses and observations provide decision-relevant evidence, and how that evidence should be combined for decision-making? To investigate and address these questions, we operationalize actionability in route-level planning over competing world hypotheses, with a limited verification budget and potentially unreliable observations. On canonical episodes, single-hypothesis commitment and cross-world voting collide in 20% and 24% of cases, respectively, while random and information-gain verification become less safe as they assimilate additional evidence. These results reveal a *Verification Paradox*: assimilating additional miscalibrated evidence can increase collision risk when correlated world errors and misaligned observations are fused overconfidently. We therefore introduce Calibration-Aware Cross-World Verification (CWV), a training-free decision layer that forms cross-world route beliefs, targets decision-critical verification, down-weights inconsistent observations through innovation-gated inverse-variance fusion, and commits using lower confidence bounds. Base CWV is collision-free across canonical camera budgets; Gated CWV removes the matched-seed high-budget negative-control failure and reduces regret by 66%. Broader evaluations identify three empirical boundaries of actionability: candidate informativeness, task sensitivity, and observation calibration. Together, these results show that generated worlds can support reliable action when decision-relevant variation and external evidence are used in a calibrated way.

1 INTRODUCTION

Generated world models increasingly serve as planning substrates for spatial agents, from satellite-to-3D street scenes and city meshes (Qian et al., 2026c; Hua et al., 2026) to large Gaussian-splat environments (Kerbl et al., 2023; Qian et al., 2026a), density-field reconstructions (Qian et al., 2023; 2026b), and action-conditioned video rollouts (DreamX Team et al., 2026). As they move from visualization to decision support, a central question is whether a generated world supports reliable downstream action. For route planning, this requires preserving task-relevant geometry so that a commitment made in the generated world remains safe in the physical one (Figure 1).

Current evaluation mainly measures visual or geometric plausibility (Qian et al., 2023; 2026c; Hua et al., 2026; Qian et al., 2026a;b), while a planner treats each generated world as an operational hypothesis. Missing or hallucinated structures, deformation, and systematic height bias can alter route clearance without destroying visual plausibility. Multiple hypotheses do not automatically solve the problem because outputs from the same model family can share structured errors (Mei et al., 2025; Bar et al., 2025). The downstream agent must therefore reason over competing hypotheses and decide which evidence should influence commitment.

This gap motivates *generated-world actionability*: whether generated worlds support reliable decisions, when their hypotheses and verification channels provide decision-relevant evidence, and how an agent should combine that evidence before commitment. We study these questions in a route-

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.

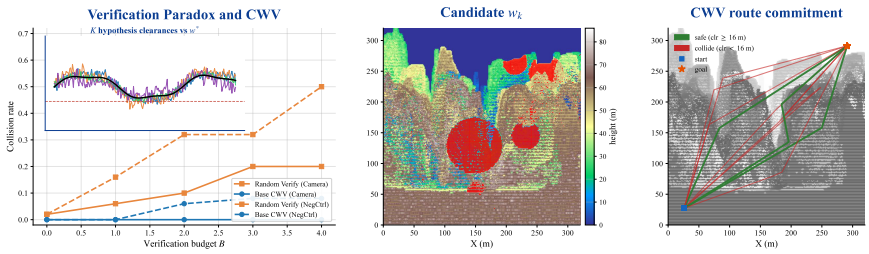

Figure 1: **Generated-world actionability and CWV.** Competing generated worlds induce different route-clearance beliefs, and naive verification can become less safe as it assimilates more evidence (left). CWV reasons over these hypotheses (center), verifies decision-critical routes, calibrates the resulting observations, and commits using a lower confidence bound (right).

level setting where a planner evaluates candidate routes over multiple world hypotheses and spends a small budget on local camera-depth verification. On canonical camera episodes, committing to a single hypothesis yields a 20% collision rate, while cross-world voting reaches 24%. Naive verification can become less safe as it consumes additional evidence: random and information-gain verification exhibit near-monotone verification-to-collision trends under both camera and adversarial negative-control settings, and random verification reaches 50% collision after four queries in negative control. We call this evidence-utilization failure the *Verification Paradox*. It arises when shared structural hallucinations make cross-world agreement unreliable and misaligned local observations are fused with excessive confidence. In this regime, the value of an observation depends not only on acquiring it, but on assigning it to the right decision and calibrating its influence before fusion (Bajcsy, 1988; Belakaria et al., 2020).

We introduce **Calibration-Aware Cross-World Verification** (CWV), a training-free decision layer that forms route beliefs across generated hypotheses, prioritizes task-critical verification, applies innovation-gated inverse-variance fusion (Kalman, 1960), and commits through a lower-confidence-bound rule (Artzner et al., 1999; Auer et al., 2002). Its stages mirror the decision interface exposed above: selection focuses the verification budget on routes that can change the decision, fusion controls how strongly new evidence moves belief, and commitment retains residual uncertainty in the final choice.

Base CWV is collision-free across canonical camera budgets. On a matched-seed negative-control sweep, Gated CWV reduces $B{=}4$ collision from 16.7% to 0% and mean regret by 66%. Broader evaluations under observation corruption, geographic transfer, cross-mesh and cross-source hypotheses, a measured LiDAR terrain reference, and an implicit video-world extension identify three empirical boundaries of actionability. Together, these results place generated-world evaluation at the decision interface rather than at appearance alone. Our main contributions are:

- We formulate generated-world actionability for route-level planning and identify the Verification Paradox, in which assimilating additional miscalibrated evidence can increase collision risk.
- We introduce CWV, a training-free decision layer combining cross-world beliefs, task-critical verification, calibration-aware inverse-variance fusion, and risk-aware commitment.
- We characterize three empirical boundaries of actionability across geographic, geometric, observational, and representational variations, with additional evidence from generated visual futures.

2 Related Work

2.1 Generated Worlds for Downstream Agents

Generated-world research is moving from view synthesis toward environments that agents can query and act within. Satellite-based reconstruction spans neural rendering, density fields, and cross-view generation (Marí et al., 2022; Qian et al., 2023; Li et al., 2024; Qian et al., 2026b); newer systems expose street-level geometry, large 3D environments, and action-conditioned rollouts (Qian et al., 2026c; Hua et al., 2026; Kerbl et al., 2023; Qian et al., 2026a; Chen et al., 2025a). Reconstruction

or rollout quality alone does not establish that downstream decisions remain valid. Existing work studies hallucination, calibration, ensemble uncertainty, and distribution shift (Mei et al., 2025; Bar et al., 2025; Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Ovadia et al., 2019), while latent-imagination methods improve the generator or policy (Ha & Schmidhuber, 2018; Hafner et al., 2020). We instead study already-generated outputs as competing hypotheses that may share task-relevant structural errors.

2.2 Belief-Space Planning and Active Perception

Belief-space planning formalizes action under uncertain state through POMDPs and scalable approximations (Kaelbling et al., 1998; Sarraute et al., 2012; Bouton et al., 2020; Mern et al., 2021b;a; Osanlou et al., 2022). Robust planning, MPC, barrier functions, and sampling-based planning provide complementary mechanisms under specified uncertainty models (Mannor et al., 2016; Duchi & Namkoong, 2021; Badings et al., 2022; Camacho & Bordons, 2013; Williams et al., 2017; Ames et al., 2019; LaValle, 2006; Karaman & Frazzoli, 2011). Active perception selects observations expected to improve downstream decisions (Bajcsy, 1988; Belakaria et al., 2020; Mern et al., 2021a); our setting challenges this expectation because world hypotheses may share correlated errors and observations may be misaligned. Confidence-bound and coherent-risk rules (Artzner et al., 1999; Rockafellar & Uryasev, 2000; Auer et al., 2002) inform our commitment rule and risk-aware baselines.

2.3 Risk-Aware Planning in 3D Representations

Recent work shows that learned 3D representations can support risk-aware planning once a scene representation is available. CATNIPS derives chance-constrained collision estimates from NeRF density (Chen et al., 2024), Splat-Nav constructs safe corridors over 3D Gaussian Splatting scenes (Chen et al., 2025b), and NeRF-Nav couples learned representations with localization and closed-loop flight (Adamkiewicz et al., 2022). Task-aware risk estimation evaluates perception failures through their downstream planning consequences (Antonante et al., 2023), while SPLANNING derives probabilistic collision bounds for planning within a Gaussian-splat scene (Michaux et al., 2025). Related work probes sensitivity to adversarial environment parameters (Hamdi et al., 2020) and uses conformal calibration to control trajectory uncertainty (Shafer & Vovk, 2008; Sarkar et al., 2022; Huang et al., 2025). These methods generally begin from one reconstructed scene, synthesized alternatives around a perception failure, or a predefined uncertainty model. We begin from K conflicting generated-world hypotheses, none assumed correct, and use their route-level disagreement to decide what to verify, how to fuse the resulting evidence, and when to commit.

3 Problem Formulation

3.1 Generated-World Route Commitment

We use a *generated world* to denote an explicit or implicit environment realization produced by a generative reconstruction system or world model and exposed to a downstream agent. In our route-level abstraction, each generated world is represented by a queryable spatial hypothesis w_k of the same physical environment. The hypotheses may differ because of reconstruction errors, hallucinated structures, or systematic model bias.

Given R candidate routes $\mathcal{R} = \{r_1, \dots, r_R\}$ and K generated-world hypotheses $\mathcal{W} = \{w_1, \dots, w_K\}$, the agent must commit to one route. Each hypothesis exposes route-level geometry, or an equivalent occupancy or clearance interface. For route r, its clearance under hypothesis w_k is

$$c_k(r) = \min_{p \in r} \operatorname{dist}(p, \operatorname{obstacles}(w_k)) \,. \tag{1}$$

A privileged reference world $w^\star$ represents the physical environment and defines the reference clearance $c^\star(r)$. Its complete geometry is hidden from the agent and used only to generate permitted local observations and to evaluate the final commitment.

3.2 Sparse Verification and Decision Objective

Before commitment, the agent may verify at most B routes. Let $\mathcal{V} \subseteq \mathcal{R}$ denote the queried routes, with $|\mathcal{V}| \leq B$. A query assigned to route r returns a partial clearance observation

$$o_r \sim q(\cdot \mid r, w^\star), \qquad \mathrm{Var}(o_r) = \sigma_o^2, \tag{2}$$

where the observation channel q need not be unbiased or correctly attributed. It may therefore represent ordinary sensor noise, corruption, occlusion, pose error, or an observation originating from a different route.

A route-commitment procedure jointly determines what to verify and which route to select:

$$\hat{r} = \pi(\mathcal{W}, \mathcal{R}, \{o_r : r \in \mathcal{V}\}), \qquad |\mathcal{V}| \leq B. \tag{3}$$

Here B is an upper bound, not a requirement to consume every available query. Because any policy feasible under budget B is also feasible under $B+1$, the Verification Paradox concerns fixed policies whose realized risk worsens when they acquire and assimilate additional evidence, rather than the availability of a larger budget itself.

A commitment is safe when its reference clearance meets the safety margin m:

$$c^\star(\hat{r}) \geq m. \tag{4}$$

Collision rate is the fraction of episodes in which this condition is violated. Throughout, "collision" denotes violation of this clearance-based safety constraint rather than literal geometric contact.

When regret is reported, we consider episodes containing at least one feasible candidate route. Let $\mathrm{cost}(r)$ denote route cost, and let

$$J^\star = \min_{r:\, c^\star(r) \geq m} \mathrm{cost}(r) \tag{5}$$

be the cost of the lowest-cost feasible route in the reference world. Regret is

$$\mathrm{regret}(\hat{r}) = \begin{cases} \mathrm{cost}(\hat{r}) - J^\star, & c^\star(\hat{r}) \geq m, \\ C_\infty - J^\star, & c^\star(\hat{r}) < m, \end{cases} \tag{6}$$

where $C_\infty = 1000$ assigns a large penalty to unsafe commitment. Average regret therefore reflects both route efficiency and the tail cost of collision.

3.3 Route-Level Actionability

We study the route-commitment layer between generated worlds and a navigation stack: the agent receives competing hypotheses and sparse local observations, then returns one route. Robot dynamics, localization, low-level control, and high-level language planning lie outside this abstraction.

For an episode distribution $\mathcal{D}$, we characterize the *route-level actionability* of a generated-world substrate under commitment procedure π by its collision-regret pair:

$$\mathcal{A}(\pi; \mathcal{D}) = \Big(\Pr_{\mathcal{D}}[c^\star(\hat{r}) < m],\ \mathbb{E}_{\mathcal{D}}[\mathrm{regret}(\hat{r})] \Big). \tag{7}$$

A substrate is more actionable when it supports a better collision-regret tradeoff under the same task specification. Actionability depends on the hypotheses, routes, safety margin, verification budget, observation channel, and commitment procedure; it is not intrinsic to a generated world. We ask whether generated worlds support reliable commitment, when their evidence is decision-relevant, and how it should be combined.

4 Method

Calibration-Aware Cross-World Verification (CWV) maps multiple generated-world hypotheses to one route commitment. Task relevance and cross-world disagreement determine where to verify; calibration determines how strongly new evidence updates belief; residual uncertainty determines commitment.

4.1 CROSS-WORLD ROUTE BELIEFS

For each route r, we compute its clearance $c_k(r)$ under every generated-world hypothesis using Equation 1. We summarize the resulting cross-world evidence by

$$\mu_r = \frac{1}{K}\sum_{k=1}^{K} c_k(r), \qquad \sigma_r^2 = \frac{1}{K}\sum_{k=1}^{K}\left(c_k(r) - \mu_r\right)^2. \tag{8}$$

We represent this evidence using a moment-matched Gaussian working belief,

$$C_r \approx \mathcal{N}\left(\mu_r, s_r^2\right), \qquad s_r^2 = \sigma_r^2 + \sigma_0^2, \tag{9}$$

where C_r denotes the unknown reference clearance. We set $\sigma_0^2 = 4\,\mathrm{m}^2$ to mitigate overconfidence when generated hypotheses agree, including from correlated errors.

4.2 TASK-CRITICAL VERIFICATION

Given a budget of B observations, CWV prioritizes routes whose verification is most likely to affect the final commitment. For each route, we define

$$\kappa_r = \mathbf{1}[\mu_r \geq m], \qquad \text{(predicted feasibility)}, \tag{10}$$

$$\beta_r = \exp\left(-\frac{(\mu_r - m)^2}{2\ell^2}\right), \qquad \text{(boundary proximity)}, \tag{11}$$

$$\alpha_r = \frac{1}{1 + \mathrm{cost}(r)/\overline{\mathrm{cost}}}, \qquad \text{(cost preference)}, \tag{12}$$

where $\overline{\mathrm{cost}}$ is the mean cost of the candidate routes. The task-criticality score is

$$I_r = \sigma_r\left(0.5 + \beta_r\right)\left(0.5 + \kappa_r\right)\left(0.5 + \alpha_r\right). \tag{13}$$

CWV verifies the top-B routes by I_r, favoring cross-world disagreement near the safety boundary at competitive cost. Unlike information gain, the score targets decision impact rather than global uncertainty reduction. We set $\ell = 8\,\mathrm{m}$ and report sensitivity in the appendix.

4.3 CALIBRATION-AWARE FUSION

A verification query on route r returns a partial clearance observation o_r with assumed variance σ_o^2. We measure its normalized disagreement with the cross-world belief using

$$\eta_r = \frac{|o_r - \mu_r|}{\sqrt{s_r^2 + \sigma_o^2}}. \tag{14}$$

CWV converts this disagreement into an effective observation variance:

$$\widetilde{\sigma}_{o,r}^2 = \begin{cases} \gamma\sigma_o^2, & \eta_r > \tau, \\ \sigma_o^2, & \eta_r \leq \tau. \end{cases} \tag{15}$$

Thus, an observation that strongly conflicts with the cross-world belief is down-weighted rather than discarded. The prior and observation precisions are

$$\pi_r^{\mathrm{prior}} = \frac{1}{s_r^2}, \qquad \pi_r^{\mathrm{obs}} = \frac{1}{\widetilde{\sigma}_{o,r}^2}. \tag{16}$$

We then apply inverse-variance fusion (Kalman, 1960):

$$\mu_r' = \frac{\pi_r^{\mathrm{prior}}\mu_r + \pi_r^{\mathrm{obs}} o_r}{\pi_r^{\mathrm{prior}} + \pi_r^{\mathrm{obs}}}, s_r'^{\,2} = \frac{1}{\pi_r^{\mathrm{prior}} + \pi_r^{\mathrm{obs}}}. \tag{17}$$

For unverified routes, we retain the prior parameters, $\mu_r' = \mu_r$ and $s_r' = s_r$. We fix $\tau = 1$ and evaluate two pre-specified gate strengths, $\gamma \in \{9, 16\}$, while holding the remaining hyperparameters fixed across experiments. The gate measures consistency between a local observation and the cross-world belief: large innovations receive less weight, while signed innovation distinguishes observation mismatch from a shared-bias correction in Section 5.4.

Table 1: Canonical Berkeley results with 24×2 observations and $n=50$. Collision columns include Wilson 95% confidence intervals. Matched-seed Gated CWV results are reported in Table 2.

	Camera $B=2$			Negctrl $B=2$			Negctrl $B=4$		
Method	Coll.	95% CI	Regret	Coll.	95% CI	Regret	Coll.	95% CI	Regret
Single World	20.0%	[11.2, 33.0]	144.9	20.0%	[11.2, 33.0]	144.9	20.0%	[11.2, 33.0]	144.9
Multiverse Vote	24.0%	[14.3, 37.4]	145.9	24.0%	[14.3, 37.4]	145.9	24.0%	[14.3, 37.4]	145.9
Random Verify	10.0%	[4.4, 21.4]	63.8	32.0%	[20.8, 45.8]	205.8	50.0%	[36.6, 63.4]	308.9
InfoGain Robust	**0.0%**	[0.0, 7.1]	39.1	14.0%	[6.9, 26.2]	115.1	14.0%	[6.9, 26.2]	114.4
TC-EMA (α=0.25)	2.0%	[0.4, 10.5]	12.5	4.0%	[1.1, 13.5]	23.6	4.0%	[1.1, 13.5]	25.9
Base CWV	**0.0%**	[0.0, 7.1]	**5.3**	6.0%	[2.1, 16.2]	64.9	8.0%	[3.2, 18.8]	84.6
Oracle	**0.0%**	[0.0, 7.1]	**0.0**	**0.0%**	[0.0, 7.1]	**0.0**	**0.0%**	[0.0, 7.1]	**0.0**

4.4 Risk-Aware Commitment and Complexity

CWV commits to the lowest-cost route whose posterior lower confidence bound clears the safety margin:

$$r^\star = \arg\min_{r:\, \mu'_r - \lambda s'_r \geq m} \mathrm{cost}(r), \qquad \lambda = 1. \tag{18}$$

If no route satisfies the constraint, CWV selects the largest $\mu'_r - \lambda s'_r$. The method without Equation 15 is *Base CWV*; the complete method is *Gated CWV*. Cross-world clearances require $O(KR)$ geometry queries, followed by $O(R \log R)$ scoring, $O(B)$ fusion, and $O(R)$ commitment. CWV requires no training; pseudocode and implementation details are in the appendix.

5 Experiments

5.1 Experimental Setup

Worlds and candidates. We evaluate on a canonical Sat3DGen-derived Berkeley reference (Qian et al., 2026c), eight geographic settings, a second Sat3DGen mesh, cross-source OSM procedural candidates (OpenStreetMap contributors, 2004), and a USGS 3DEP LiDAR-measured bare-earth terrain reference (U.S. Geological Survey, 2023). The canonical setting contains $K=5$ structurally perturbed hypotheses with shared and hypothesis-specific geometric errors, whereas the cross-mesh and cross-source constructions do not share the reference vertex layout. Additional details are provided in the appendix.

Observations and methods. Camera verification returns partial route-clearance evidence from depth images rendered using the `trimesh` ray-mesh backend (Dawson-Haggerty et al., 2019). Negative control assigns an observation from a different route, while the corruption setting replaces aligned observations with random-route observations with probability p; the agent never observes reference clearance directly. We compare Single World, Multiverse Vote, Random Verify, InfoGain Verify, InfoGain Robust, TC-EMA, Base CWV, Gated CWV, and Oracle.

Metrics and statistics. We report collision rate under the $m=16\,\mathrm{m}$ safety margin and regret from Equation 6. Collision proportions use Wilson 95% confidence intervals (Newcombe, 1998), and regret uses 2000-sample bootstrap intervals (Efron & Tibshirani, 1993). Paired comparisons use McNemar and Wilcoxon tests (McNemar, 1947; Wilcoxon, 1945), while monotone budget-to-collision trends use Spearman correlation with exact permutation tests (Spearman, 1904). Details are provided in the appendix.

5.2 Direct Use and the Verification Paradox

Direct use of generated worlds and naive verification both fail on canonical Berkeley. Single World collides in 20% of camera episodes and Multiverse Vote in 24%, showing that correlated candidate errors can make agreement misleading. Random Verify rises from 10% collision at $B=2$ to 20% at $B=4$ under camera and from 32% to 50% under negative control (Table 1). Its budget trend has Spearman $\rho=0.975$ in both modes ($p=0.017$), while InfoGain Verify reaches $\rho=1.000$ ($p=0.008$).

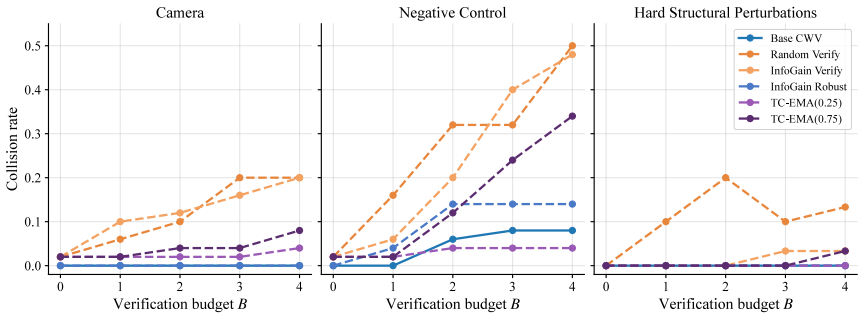

Figure 2: **Verification Paradox.** Collision rate versus verification budget under camera observations (left), adversarial negative control (middle), and hard structural perturbations (right). Random Verify and InfoGain Verify worsen as additional queries are assimilated, whereas Base CWV remains near zero.

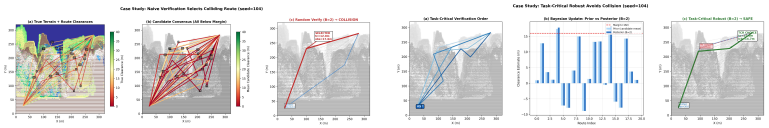

Figure 3: **Case-level behavior on canonical Berkeley.** Random Verify misattributes a $17.6\,\mathrm{m}$ observation to a route with $12.0\,\mathrm{m}$ reference clearance and violates the safety margin (left). Base CWV instead selects a safe route with $25.0\,\mathrm{m}$ reference clearance (right).

The experiments localize the paradox to two decisions: which route is queried and how its observation updates belief. Random selection spends budget on routes with little decision impact, while hard override lets a misaligned observation replace the cross-world prior. Repeating either mistake across additional queries produces the budget-dependent degradation. The same pattern appears across geographic, cross-mesh, cross-source, LiDAR, and implicit video-world evaluations (Sections 5.5–5.6). Figure 3 shows the mechanism at the episode level.

5.3 Base CWV and Component Ablations

Table 1 contains the primary comparison. Base CWV achieves 0% collision at every canonical camera budget. At $B{=}2$, it matches InfoGain Robust on safety while reducing mean regret from 39.1 to 5.3, showing that task-critical selection improves efficiency beyond robust fusion alone. The negative-control comparison motivates the calibration gate in Section 5.4.

The ablation in Table 2 shows that Bayesian fusion and LCB commitment are complementary: removing either introduces collisions, while removing both produces the largest safety and regret degradation. Selection also matters independently of fusion, as Random Robust remains collision-free but has higher $B{=}2$ regret than Base CWV. Figure 4 summarizes the safety-efficiency tradeoff: Base CWV concentrates near the low-collision, low-regret corner while naive verification climbs to the upper right and conservative baselines shift rightward at low collision. TC-EMA is more resistant to pure negative control at large budgets because of its slower update, but sacrifices clean-camera efficiency; Gated CWV addresses this tradeoff in Section 5.4. Additional paired tests, selection ablations, and regret distributions are provided in the appendix.

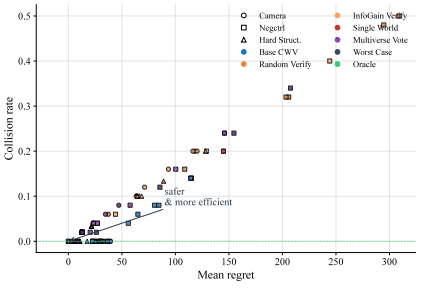

Figure 4: **Safety-efficiency tradeoff.** Collision rate versus mean across camera, control, perturbation experiments.

Table 2: **Calibration and component analysis.** Top: Gated CWV under negative control across verification budgets. Bottom: fusion × commitment ablation on camera, where each entry at $B{=}2, 4$ reports collision rate / mean regret.

Setting	Method	$B{=}0$	$B{=}1$	$B{=}2$	$B{=}3$	$B{=}4$
Negctrl	Base CWV (Coll.%)	0.0	0.0	3.3	10.0	16.7
	Gated $\gamma{=}16$ (Coll.%)	0.0	0.0	0.0	0.0	0.0
	Gated $\gamma{=}9$ (Coll.%)	0.0	0.0	0.0	3.3	6.7
	Base CWV (Regret)	22.5	35.0	38.6	82.4	146.3
	Gated $\gamma{=}16$ (Regret)	22.5	25.5	25.5	44.7	49.3
	Gated $\gamma{=}9$ (Regret)	22.5	30.0	38.6	61.2	75.9
Camera ablation	Bayesian + LCB	–	–	0.0% / 5.1	–	0.0% / 29.7
	Bayesian + Mean	–	–	0.0% / 0.0	–	3.3% / 18.9
	Hard + LCB	–	–	0.0% / 30.2	–	3.3% / 88.8
	Hard + Mean	–	–	6.7% / 38.9	–	20.0% / 115.4

Table 3: Multi-region generalization across eight geographic settings. CWV_B denotes Base CWV camera collision at budget B, and CWV-NC_4 its negative-control collision at $B{=}4$. $n{=}30$ per mode except Chicago Lakeview negative control ($n{=}53$).

Region	Single	Random_4	CWV_0	CWV_2	CWV_4	CWV-NC_4	n (cam / nc)
Berkeley OSM	10.0%	40.0%	0.0%	0.0%	0.0%	6.7%	30 / 30
Berkeley Sat3DGen	10.0%	13.3%	0.0%	0.0%	0.0%	16.7%	30 / 30
Seattle Capitol Hill	0.0%	3.3%	0.0%	0.0%	0.0%	0.0%	30 / 30
Chicago Lakeview	16.7%	10.0%	0.0%	0.0%	0.0%	7.5%	30 / 53
Chicago Loop	20.0%	20.0%	3.3%	3.3%	3.3%	10.0%	30 / 30
Manhattan Midtown	26.7%	43.3%	0.0%	10.0%	20.0%	10.0%	30 / 30
Phoenix Suburb	23.3%	33.3%	3.3%	6.7%	10.0%	20.0%	30 / 30
Seattle Downtown	30.0%	33.3%	0.0%	3.3%	10.0%	16.7%	30 / 30

5.4 Calibration Diagnosis and Gated CWV

Negative control isolates the effect of observation misalignment: at $B{=}4$, Base CWV records 8.0% collision in the canonical $n{=}50$ run and 16.7% in the matched-seed $n{=}30$ sweep. Higher depth resolution reduces the latter to 10.5%. The calibration audit shows conservative LCB thresholding (0.6% false certification) while posterior uncertainty continues to tighten with additional evidence, motivating an explicit consistency correction. Full diagnostics are in the appendix.

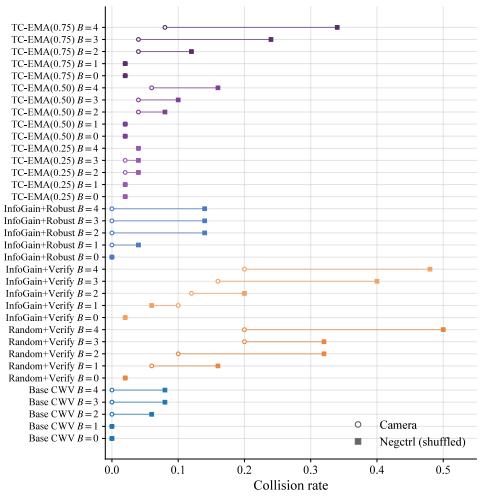

Figure 5: **Observation sensitivity.** Canonical camera and negative-control diagnostics.

The innovation gate inflates variance for large innovations and fires on roughly 20–30% of negative-control queries. At $B{=}4$, $\gamma{=}16$ reduces collision from 16.7% to 0% and mean regret from 146.3 to 49.3 without changing camera performance (Table 2). The effect transfers to corruption, three of four out-of-Berkeley evaluations, and held-out seeds. Chicago Loop represents the shared-bias regime, where signed innovation separates corrective low-side evidence from ordinary mismatch; the diagnostic recovers 0% collision at $B{=}2$. Full calibration results are in the appendix. On this matched-seed sweep, gating changes the verification trend itself: Base CWV rises from 0% to 16.7%

collision as additional observations are assimilated, whereas Gated CWV remains at 0% across the same budgets.

5.5 Generalization and Actionability Boundaries

Across eight regions, the budget sensitivity of naive verification persists, while Base CWV reduces collision and Gated CWV further improves three of four targeted out-of-Berkeley stress tests together with a held-out seed range (Table 3). Chicago Loop exhibits the shared-bias regime analyzed through signed innovation in Section 5.4.

Three empirical boundaries emerge. *Candidate informativeness*: when Elevate3D (krdgomer, 2024) is nearly uncorrelated with the reference ($|\rho| < 0.15$), additional verification carries little useful world signal. *Task sensitivity*: the altitude sweep separates regimes with no feasible route from regimes in which all feasible routes are already safe. *Observation calibration*: higher depth fidelity reduces the negative-control gap, while variance correction closes the remaining matched-seed gap. Figure 6 shows a cross-source case where task-critical selection matches Oracle.

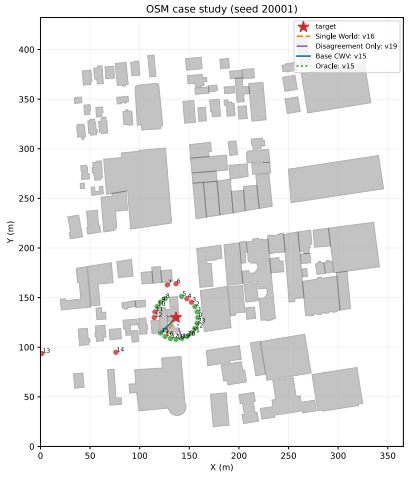

Figure 6: **Cross-source target selection.** On OSM candidates, task-critical selection matches the Oracle target.

5.6 External Validity

Beyond canonical Sat3DGen, the same decision principle transfers to LiDAR terrain, cross-source OSM candidates, cross-mesh Sat3DGen candidates, and implicit video rollouts. On LiDAR, naive high-budget verification appears in the regret tail; on OSM, Base CWV remains collision-free despite a $-25.4\,\mathrm{m}$ systematic height gap; DreamX (DreamX Team et al., 2026) exposes growing cross-seed disagreement, whereas Sat2Density++ (Qian et al., 2026b) provides a low-variation positive control. Appendix H extends the same actionability question to generated visual futures, where previewed consequences are checked against production execution.

6 Discussion

Empirical Boundaries of Actionability. Across the evaluated settings, actionability depends on whether candidate variation is task-relevant, whether the task exposes meaningful differences among actions, and whether verification is calibrated well enough to update belief.

Implications for Generated-World Planning. Generated worlds are most useful as task-conditioned hypotheses rather than as fixed maps. What matters is whether their variation supports the downstream decision and whether new evidence predicts realized consequences. The creative-agent study in Appendix H reaches the same conclusion for visual futures: present quality, future editability, and execution calibration need not rank candidates in the same order.

7 Conclusion

Generated worlds can support planning when their uncertainty is carried into the decision rather than hidden behind a plausible output. We formulate generated-world actionability and identify the Verification Paradox, then use CWV to connect task-critical verification, calibrated evidence fusion, and uncertainty-aware commitment. Base CWV is collision-free across canonical camera budgets, and Gated CWV removes the matched-seed high-budget negative-control failure.

AI USE STATEMENT

Generative AI tools were used for literature retrieval and discovery, research ideation, code navigation and implementation assistance, and language editing. All AI-assisted outputs were reviewed and verified by the authors, who take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies decision-making over generated environments using public datasets, generated world representations, and simulated planning settings. It does not involve human subjects, private or sensitive personal data, or deployment in real-world safety-critical systems. The experiments are intended to study reliability and calibration of generated-world planning rather than to support direct autonomous deployment.

REPRODUCIBILITY STATEMENT

The generated-world actionability formulation and CWV method are specified in Sections 3 and 4. Experimental settings, evaluation metrics, and main baselines are described in Section 5.1. The appendix provides the full algorithm, statistical methodology, calibration diagnostics, sensitivity analyses, geographic and representation extensions, and the generated-future creative-agent experiments, together with the corresponding hyperparameters and evaluation settings.

REFERENCES

Michał Adamkiewicz, Timothy Chen, Adam Caccavale, Rachel Gardner, Preston Culbertson, Jeannette Bohg, and Mac Schwager. Vision-only robot navigation in a neural radiance world. *IEEE Robotics and Automation Letters*, 7(2):4606–4613, 2022.

Aaron D. Ames, Samuel Coogan, Magnus Egerstedt, Gennaro Notomista, Koushil Sreenath, and Paulo Tabuada. Control barrier functions: Theory and applications. In *Proceedings of the 18th European Control Conference (ECC)*, pp. 3420–3431, 2019.

Pasquale Antonante, Sushant Veer, Karen Leung, Xinshuo Weng, Luca Carlone, and Marco Pavone. Task-aware risk estimation of perception failures for autonomous vehicles. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.100.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002.

Thom S. Badings, Alessandro Abate, Nils Jansen, David Parker, Hasan A. Poonawala, and Marielle Stoelinga. Sampling-based robust control of autonomous systems with non-Gaussian noise. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022. Distinguished Paper.

Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988.

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

Syrine Belakaria, Aryan Deshwal, and Janardhan Rao Doppa. Multi-fidelity multi-objective Bayesian optimization: An output space entropy search approach. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

Maxime Bouton, Jana Tumova, and Mykel J. Kochenderfer. Point-based methods for model checking in partially observable Markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

Eduardo F. Camacho and Carlos Bordons. *Model Predictive Control*. Springer, 2013.

Anthony Chen, Wenzhao Zheng, Yida Wang, Xueyang Zhang, Kun Zhan, Peng Jia, Kurt Keutzer, and Shanghang Zhang. GeoDrive: 3D geometry-informed driving world model with precise action control. *arXiv preprint arXiv:2505.22421*, 2025a.

Timothy Chen, Preston Culbertson, and Mac Schwager. CATNIPS: Collision avoidance through neural implicit probabilistic scenes. *IEEE Transactions on Robotics*, 40:2712–2728, 2024.

Timothy Chen, Ola Shorinwa, Joseph Bruno, Aiden Swann, Javier Yu, Weijia Zeng, Keiko Nagami, Philip Dames, and Mac Schwager. Splat-Nav: Safe real-time robot navigation in Gaussian splatting maps. *IEEE Transactions on Robotics*, 41:2765–2784, 2025b.

Michael Dawson-Haggerty et al. trimesh: Pythonic library for triangular meshes. `https://trimesh.org/`, 2019.

DreamX Team, Yancheng Bai, Rui Chen, Xiangxiang Chu, Rujing Dang, Hao Dou, Bingjie Gao, Qiwen Gu, Siyu Hong, Jiachen Lei, et al. Dreamx-world 1.0: A general-purpose interactive world model. *arXiv preprint arXiv:2606.16993*, 2026.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *Annals of Statistics*, 49(3):1378–1406, 2021.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations (ICLR)*, 2020.

Abdullah Hamdi, Matthias Müller, and Bernard Ghanem. SADA: Semantic adversarial diagnostic attacks for autonomous applications. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968.

Tongyan Hua, Dongli Wu, Jinjing Zhu, Yinrui Ren, Zhongcheng Hong, Ying-Cong Chen, Hui Xiong, and Wufan Zhao. Sat2City v2: Native 3D city asset generation from a single satellite image. *arXiv preprint arXiv:2606.24138*, 2026.

Huiqun Huang, Sihong He, and Fei Miao. CUQDS: Conformal uncertainty quantification under distribution shift for trajectory prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

Rudolf E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

Sertac Karaman and Emilio Frazzoli. Sampling-based algorithms for optimal motion planning. *International Journal of Robotics Research*, 30(7):846–894, 2011.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023.

krdgomer. Elevate3D: Deep learning-powered 3D city reconstruction from satellite imagery. https://github.com/krdgomer/Elevate3D, 2024. Open-source software, MIT license.

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

Steven M. LaValle. *Planning Algorithms*. Cambridge University Press, 2006.

Zuoyue Li, Zhenqiang Li, Zhaopeng Cui, Marc Pollefeys, and Martin R. Oswald. Sat2Scene: 3D urban scene generation from satellite images with diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. *ACM SIGGRAPH Computer Graphics*, 21(4):163–169, 1987.

Shie Mannor, Ofir Mebel, and Huan Xu. Robust MDPs with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509, 2016.

Roger Marí, Gabriele Facciolo, and Thibaud Ehret. Sat-NeRF: Learning multi-view satellite photogrammetry with transient objects and shadow modeling using RPC cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

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.

John Mern, Anil Yildiz, Lawrence Bush, Tapan Mukerji, and Mykel J. Kochenderfer. Improved POMDP tree search planning with prioritized action branching. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pp. 11888–11894, 2021a.

John Mern, Anil Yildiz, Zachary Sunberg, Tapan Mukerji, and Mykel J. Kochenderfer. Bayesian optimized Monte Carlo planning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pp. 11880–11887, 2021b.

Jonathan B. Michaux, Seth Isaacson, Challen Enninful Adu, Adam Li, Rahul Kashyap Swayampakula, Parker Ewen, Sean Rice, Katherine A. Skinner, and Ram Vasudevan. Let us make a splan: Risk-aware trajectory optimization in a normalized gaussian splat. *IEEE Transactions on Robotics*, 41:4380–4397, 2025. doi: 10.1109/TRO.2025.3584559.

Robert G. Newcombe. Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine*, 17(8):857–872, 1998.

OpenStreetMap contributors. Openstreetmap. https://www.openstreetmap.org/, 2004. Open Database License.

Kevin Osanlou, Jeremy Frank, Andrei Bursuc, Tristan Cazenave, Eric Jacopin, Christophe Guettier, and J. Benton. Solving disjunctive temporal networks with uncertainty under restricted time-based controllability using tree search and graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Ming Qian, Jincheng Xiong, Gui-Song Xia, and Nan Xue. Sat2Density: Faithful density learning from satellite-ground image pairs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Ming Qian, Tianjian Ouyang, Mingchao Sun, Zijian Wang, et al. ABot-Earth 0.5: Generative 3D Earth model. *arXiv preprint arXiv:2606.09967*, 2026a.

Ming Qian, Bin Tan, Qiuyu Wang, Xianwei Zheng, Hanjiang Xiong, Gui-Song Xia, Yujun Shen, and Nan Xue. Seeing through satellite images at street views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(5):5692–5709, 2026b. doi: 10.1109/TPAMI.2026.3652860.

Ming Qian, Zimin Xia, Changkun Liu, Shuailei Ma, Wen Wang, Zeran Ke, Bin Tan, Hang Zhang, and Gui-Song Xia. Sat3DGen: Comprehensive street-level 3D scene generation from single satellite image. In *International Conference on Learning Representations (ICLR)*, 2026c.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.

Atrisha Sarkar, Kate Larson, and Krzysztof Czarnecki. Generalized dynamic cognitive hierarchy models for strategic driving behavior. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

Carlos Sarraute, Olivier Buffet, and Jörg Hoffmann. POMDPs make better hackers: Accounting for uncertainty in penetration testing. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2012.

Will Schroeder, Ken Martin, and Bill Lorensen. *The Visualization Toolkit: An Object-Oriented Approach to 3D Graphics*. Kitware, 4 edition, 2006.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Charles Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101, 1904.

U.S. Geological Survey. 3D elevation program (3DEP). `https://www.usgs.gov/3d-elevation-program`, 2023.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

Grady Williams, Andrew Aldrich, and Evangelos A. Theodorou. Model predictive path integral control: From theory to parallel computation. *Journal of Guidance, Control, and Dynamics*, 40 (2):344–357, 2017. doi: 10.2514/1.G001921.

A Additional Experimental Setup

A.1 Notation

This appendix provides supplementary experiments, implementation details, and diagnostics for the main paper. We follow the notation of the main paper.

Symbol	Meaning	Default
K	Number of generated hypotheses	5
R	Number of candidate routes	7 or 20
B	Verification budget	$\{0, 1, 2, 3, 4\}$
m	Safety margin	16 m
μ_r	Cross-world mean clearance on r	—
σ_r	Cross-world std of clearance on r	—
σ_0^2	Prior variance floor	4
σ_o	Assumed observation std	8 m
τ	Innovation gate threshold	1.0
γ	Gated variance-inflation factor	$\{9, 16\}$
λ	LCB coefficient	1.0
C_∞	Collision penalty in regret	1000
ℓ	Task-critical bell width	8 m

Table S4: Notation and canonical defaults. $R{=}7$ for discrete-route and multi-region settings, $R{=}20$ for continuous-mesh evaluations, and DreamX uses seven commitment depths.

Computing Environment. Experiments were conducted on NVIDIA H200 GPUs with 140 GB memory per GPU. We used AI code assistants for code navigation, implementation assistance, and language editing.

A.2 Reference Worlds and Candidate Construction

Reference worlds. The canonical reference is a Sat3DGen-derived Berkeley mesh with 274,564 vertices, produced from a single satellite image using Sat3DGen (Qian et al., 2026c). Cross-mesh experiments use a second Sat3DGen output with 279,186 vertices generated from a different Berkeley satellite input. Cross-source experiments use OpenStreetMap-derived procedural meshes rasterized into the Sat3DGen coordinate frame (OpenStreetMap contributors, 2004). Multi-region experiments cover Berkeley OSM, Berkeley Sat3DGen, Chicago Loop, Chicago Lakeview, Manhattan Midtown, Phoenix Suburb, Seattle Downtown, and Seattle Capitol Hill. The physical-terrain experiment uses a USGS 3DEP LiDAR-measured bare-earth digital terrain model over Alameda County (U.S. Geological Survey, 2023), resampled to $4\,\mathrm{m}$ over a $320\,\mathrm{m} \times 320\,\mathrm{m}$ downtown-Berkeley tile. Route clearance is computed by geometric intersection or nearest-surface queries for mesh backends and by interpolation along the route for height-field backends.

Canonical candidates. Each canonical episode contains $K{=}5$ structurally perturbed hypotheses generated from the reference mesh using five error families: regional systematic height offsets of up to $\pm 15\,\mathrm{m}$ over Gaussian-weighted regions covering 20–40% of the map; missing structures implemented by reducing heights within 2–4 regions to 10–30% of their reference values; hallucinated obstacles implemented by raising 2–4 regions to 60–90% of the maximum reference height; a shared Gaussian-weighted bias of up to $\pm 20\,\mathrm{m}$ inherited by all candidates; and per-vertex Gaussian noise with $\sigma{=}3\,\mathrm{m}$. These candidates share the reference vertex layout and are therefore described as structurally perturbed rather than fully independent. Cross-mesh experiments instead use two real Sat3DGen outputs, while cross-source experiments use OSM procedural candidates against a Sat3DGen reference; neither construction shares the canonical reference vertex layout.

A.3 Observation Protocol

Camera observations. Each verification query renders partial depth evidence from the reference geometry using the `trimesh` ray-mesh intersector (Dawson-Haggerty et al., 2019). The canonical

Algorithm S1 Calibration-Aware Cross-World Verification

Input: Generated hypotheses $\{w_k\}_{k=1}^{K}$, candidate routes $\{r_j\}_{j=1}^{R}$, verification budget B, safety margin m, gate parameters (τ, γ), LCB coefficient λ, prior floor σ_0, obs. std σ_o
Output: Route commitment $r^\star$

For each route r, compute clearances $\{c_k(r)\}_{k=1}^{K}$
Compute μ_r, σ_r; set $s_r^2 \leftarrow \sigma_r^2 + \sigma_0^2$
Compute task-critical scores I_r
Let $\mathcal{V} \leftarrow$ top-B routes by I_r
for each $r \in \mathcal{V}$ **do**
    Acquire a local observation o_r of route r
    Compute innovation η_r
    if $\eta_r > \tau$ **then**
        $\widetilde{\sigma}_{o,r}^2 \leftarrow \gamma\, \sigma_o^2$ (gate triggered)
    else
        $\widetilde{\sigma}_{o,r}^2 \leftarrow \sigma_o^2$
    end if
    Update $\mu'_r, {s'_r}^2$ using $\widetilde{\sigma}_{o,r}^2$
end for
For $r \notin \mathcal{V}$, set $\mu'_r \leftarrow \mu_r,\ {s'_r}^2 \leftarrow s_r^2$ (unverified)
Compute LCB $L_r \leftarrow \mu'_r - \lambda s'_r$
$\mathcal{F} \leftarrow \{r : L_r \geq m\}$
if $\mathcal{F} \neq \emptyset$ **then**
    return $r^\star \leftarrow \arg\min_{r \in \mathcal{F}} \mathrm{cost}(r)$
else
    return $r^\star \leftarrow \arg\max_r L_r$
end if

configuration uses 24-pixel depth resolution and two camera poses per query, denoted 24×2, while the observation-quality experiment additionally evaluates a 48×3 configuration. Pixel depths receive Gaussian noise with standard deviation $0.5\,\mathrm{m}$. Visible route segments are back-projected into a partial point cloud, while unobserved segments retain their prior beliefs. The resulting route-level observation and assumed variance are passed to the calibration-aware fusion rule described in the main paper. Reference clearance is used only for observation generation and final evaluation and is never exposed directly to the commitment procedure.

Negative control and corruption. For a query assigned to route r, negative-control mode returns an observation generated from a randomly selected route $r' \neq r$, modeling severe route-attribution error caused by pose mislocalization, occlusion, or observation-to-route mismatch. Corruption mode independently replaces an aligned camera observation with a random-route observation with probability p. Compared strategies within an experiment receive the same reference world, candidate hypotheses, route set, and episode seeds.

A.4 Method Algorithm

We present the algorithm as Algorithm S1.

B Full Canonical Grid and Statistical Methodology

B.1 Strategy Definitions and Full Grid

Single World commits using the first generated hypothesis. *Multiverse Vote* selects the route preferred by the majority of hypotheses. *Random Verify* queries B randomly selected routes and hard-overrides each queried belief with its observation. *InfoGain Verify* queries routes with the largest cross-world clearance standard deviation and applies the same hard override. *InfoGain Robust* uses the same selection rule with inverse-variance Bayesian fusion and LCB commitment. *TC-EMA* uses task-critical selection and an exponential moving-average update. *Base CWV* combines task-

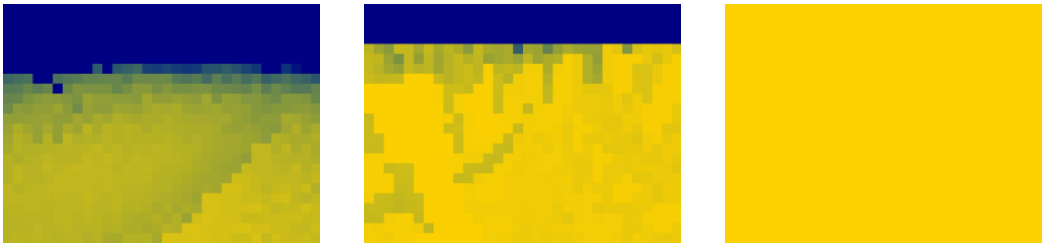

Figure S7: Depth-camera observation across three candidate hypotheses of the same Berkeley scene. Left: under-estimated building heights. Middle: approximately correct heights. Right: hallucinated tall structure. The same camera pose yields three different depth images and thus three route-clearance estimates.

Strategy (camera)	B=0	B=1	B=2	B=3	B=4
oracle	0.0	—	—	—	—
single_world	20.0	—	—	—	—
average_world	2.0	—	—	—	—
worst_case	0.0	—	—	—	—
cvar_01	0.0	—	—	—	—
cvar_02	0.0	—	—	—	—
multiverse_vote	24.0	—	—	—	—
random_verify	2.0	6.0	10.0	20.0	20.0
infogain_verify	2.0	10.0	12.0	16.0	20.0
infogain_robust	0.0	0.0	0.0	0.0	0.0
tc_ema_025	2.0	2.0	2.0	4.0	4.0
tc_ema_050	2.0	2.0	4.0	6.0	6.0
tc_ema_075	2.0	2.0	4.0	6.0	8.0
task_critical_robust	0.0	0.0	0.0	0.0	0.0

Table S5: Full camera-mode collision % grid on canonical Berkeley, n=50. — indicates a budget-independent strategy.

critical selection, inverse-variance Bayesian fusion, and LCB commitment without the innovation gate. *Gated CWV* additionally applies the innovation gate. *Oracle* selects the minimum-cost route satisfying the reference-clearance constraint and is used only as an upper-bound comparison.

The full grid contains 18 base strategies and 18 gated variants. The base strategies are grouped as follows.

No verification: Single World, Average World, Worst Case, Multiverse Vote, CVaR-0.1, CVaR-0.2, and Oracle.

Hard-override verification: Random Verify, Disagreement Verify, InfoGain Verify, and Task-Critical Verify.

Robust verification: Random Robust, Disagreement Robust, InfoGain Robust, and Base CWV.

Exponential moving average: TC-EMA with $\alpha \in \{0.25, 0.50, 0.75\}$.

Gated CWV variants: (τ, γ) swept over $\tau \in \{1.0, 1.5, 2.0, 2.5, 3.0, 4.0\}$ and $\gamma \in \{4, 9, 16\}$ ($6 \times 3 = 18$ combinations); the two operating points reported in the main paper are $(1.0, 16)$ and $(1.0, 9)$.

Tables S5 and S6 report the complete camera and negative-control collision grids. Figures S8 and S9 summarize the maximum-budget comparison and the full base-strategy grid.

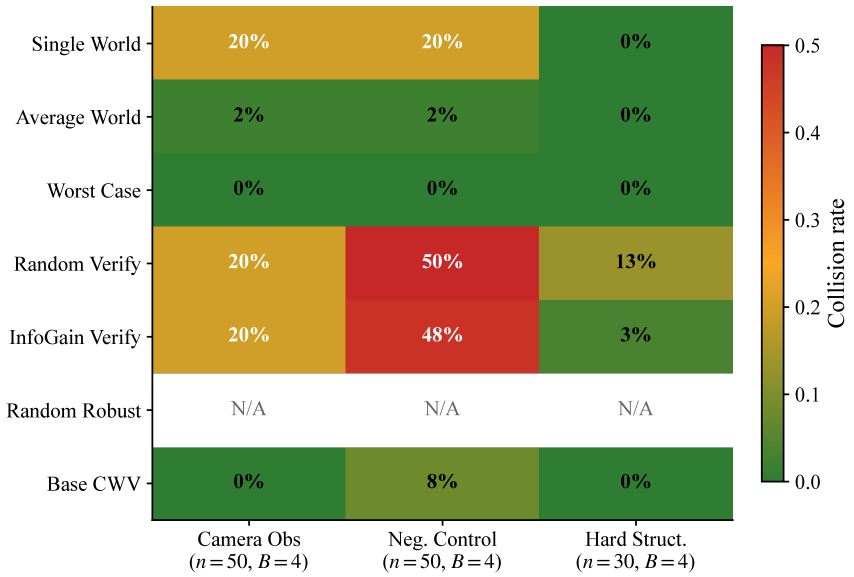

Figure S8: Cross-experiment collision rate at maximum budget (B=4) for a curated method set. Base CWV remains at 0% under canonical camera and hard structural perturbations and reaches 8% under negative control; hard-override strategies degrade most under adversarial observations.

Strategy (negctrl)	B=0	B=1	B=2	B=3	B=4
task_critical_robust	0.0	0.0	6.0	8.0	8.0
random_verify	2.0	16.0	32.0	32.0	50.0
infogain_verify	2.0	6.0	20.0	40.0	48.0
infogain_robust	0.0	4.0	14.0	14.0	14.0
tc_ema_025	2.0	4.0	4.0	4.0	4.0
tc_ema_050	2.0	4.0	8.0	12.0	16.0
tc_ema_075	2.0	8.0	20.0	28.0	34.0

Table S6: Full negctrl-mode collision % grid on canonical Berkeley, n=50.

B.2 Statistical Methodology

B.2.1 Confidence Intervals.

All collision proportions in the main paper carry Wilson score 95% confidence intervals (Newcombe, 1998). Wilson intervals have superior coverage for small samples and boundary proportions compared to Wald intervals. For n=50, an observed 0% collision produces the interval $[0\%, 7.1\%]$; an observed 20% collision produces $[11.2\%, 33.0\%]$.

B.2.2 Paired Tests.

Every episode presents the same reference, hypotheses, and routes to every strategy. We use McNemar's test with continuity correction for binary collision comparisons (McNemar, 1947) and Wilcoxon signed-rank for continuous regret comparisons (Wilcoxon, 1945). Under Bonferroni correction ($\alpha_{\text{corr}} = 0.05/74$), 10 of the 74 pairwise tests remain significant.

B.2.3 Bootstrap Regret CI.

Regret confidence intervals use 2000 bootstrap resamples with replacement, reporting the 2.5% and 97.5% percentile bounds (Efron & Tibshirani, 1993).

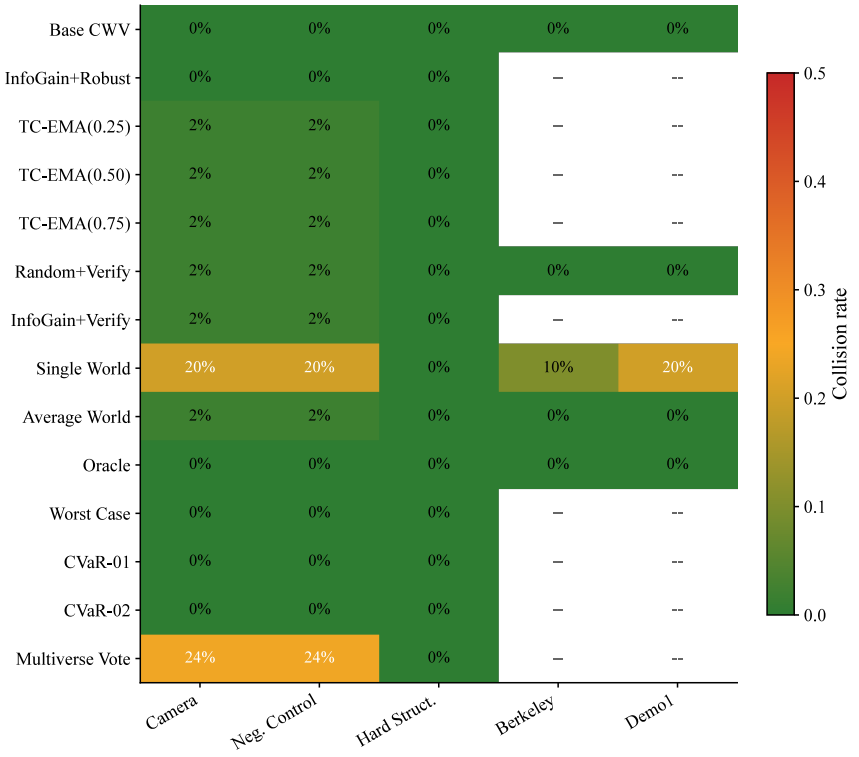

Figure S9: Strategy heatmap over the full base-strategy grid on canonical Berkeley, n=50. Cool colors indicate low collision; the paradox regions are the warm bands in the top block. This appendix figure complements Figure S8 by including every base strategy at its best budget rather than the fixed B=4 contrast.

Comparison	McNemar χ^2	p	mean Δreg.
CWV vs. single	8.10	0.0044	−122.4
CWV vs. vote	9.09	0.0025	−140.6
CWV vs. random	5.14	0.0234	−58.5
CWV vs. infogain	4.17	0.0411	−66.0
CWV vs. worst	0.00	1.0000	−31.2
CWV vs. $\text{CVaR}_{0.1}$	0.00	1.0000	−26.2

Table S7: Paired collision tests and mean regret differences on Berkeley camera at B=2, with n=50. Negative Δ regret favors Base CWV.

B.2.4 Spearman ρ for the Paradox.

We compute Spearman ρ over the five budgets $B \in \{0, 1, 2, 3, 4\}$ and obtain exact one-sided p-values by enumerating all $5! = 120$ rank permutations. Random Verify yields $\rho = 0.975$ and $p = 0.017$, while InfoGain Verify yields $\rho = 1.000$ and $p = 0.008$, the smallest attainable value under this test. Table S8 reports the complete results.

C Canonical Case Studies and Aggregate Views

We provide additional canonical ablations and aggregate views of the Verification Paradox.

Strategy	Camera		Negctrl	
	ρ	p	ρ	p
Random Verify	$+0.975$	0.017	$+0.975$	0.017
InfoGain Verify	$+1.000$	0.008	$+1.000$	0.008
Disagreement Verify	$+0.900$	0.042	$+0.900$	0.042
TC-EMA(α=0.75)	$+0.900$	0.042	$+1.000$	0.008
InfoGain Robust	-0.447	—	$+0.894$	0.050
Base CWV	$\equiv 0$	—	$+0.949$	0.033
Gated CWV	$\equiv 0$	—	$\equiv 0$	—

Table S8: Verification Paradox scores. Spearman ρ over $B \in \{0, 1, 2, 3, 4\}$ for collision rate, with exact one-sided permutation p-values from the 5! rank-permutation null distribution. Positive ρ indicates monotone worsening with budget. Rows with $\rho \equiv 0$ have zero collision at every budget. n=50 per (strategy, mode).

Strategy	Coll. all	Regret B=2	Regret B=4
Random Robust	0%	32.8	40.2
Disagreement Rob.	0%	29.4	38.5
InfoGain Robust	0%	39.1	33.7
Base CWV	0%	**5.3**	26.5

Table S9: Selection ablation, Berkeley camera. All strategies use Bayesian fusion and LCB commit; only the verification-target selector differs. Task-critical selection reduces regret by an order of magnitude at B=2. n=50.

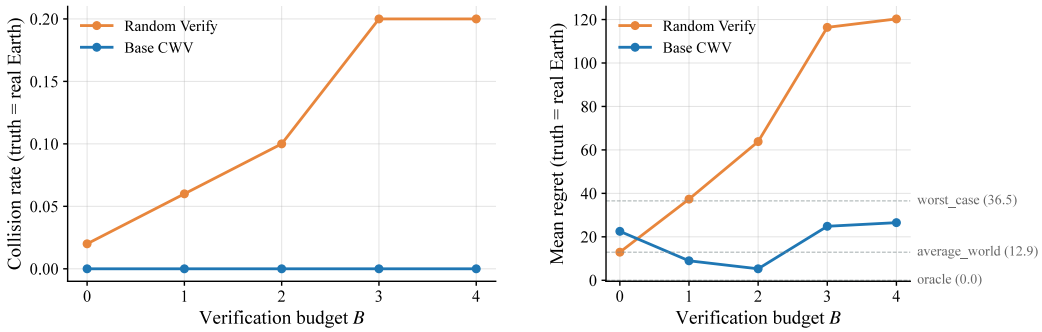

Figure S10: Reference-grounded curves on the Sat3DGen Berkeley reference. Left: collision rate vs. budget. Right: mean regret vs. budget. Base CWV holds at zero collision at every budget while its regret decreases with B; naive verification worsens with B.

D Calibration and Gate Diagnostics

We examine whether residual negative-control failures arise from posterior calibration and evaluate the innovation gate across corruption and geographic shifts.

We first distinguish failures of the final commitment rule from failures of posterior calibration. The low LCB-violation rate in Table S10 points to the latter: inconsistent observations reduce posterior uncertainty without reliably improving clearance estimates. We then evaluate whether variance inflation transfers across corruption, geographic shifts, and shared-bias cases.

E Actionability Boundaries

We evaluate three boundaries of actionability: candidate informativeness, task tightness, and observation fidelity.

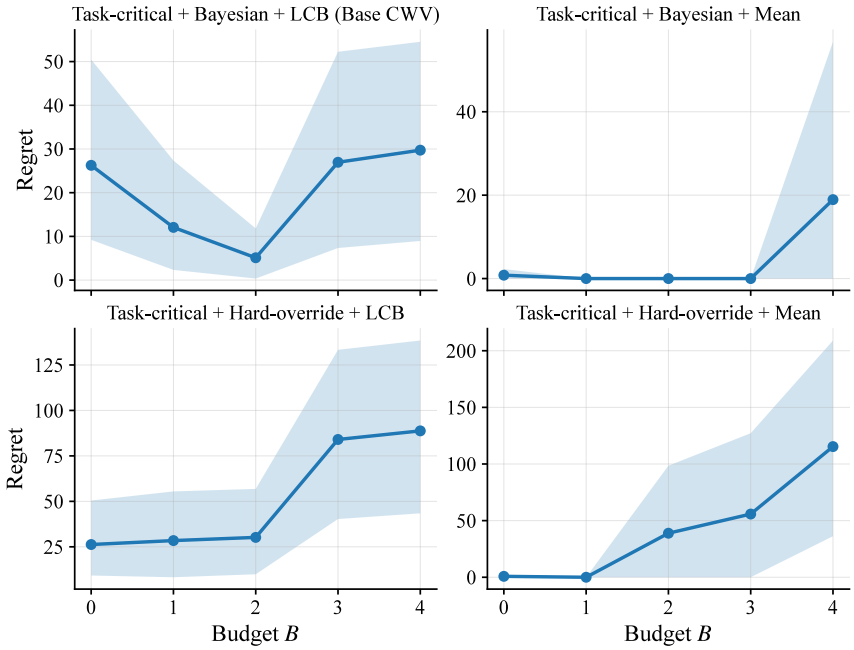

Figure S11: 2×2 ablation grid on Berkeley camera. Rows: Bayesian vs. hard-override fusion. Columns: LCB vs. mean commit. Collision at $B{=}4$ is only zero when both LCB and Bayesian fusion are present; mean commit minimizes regret only when combined with Bayesian fusion.

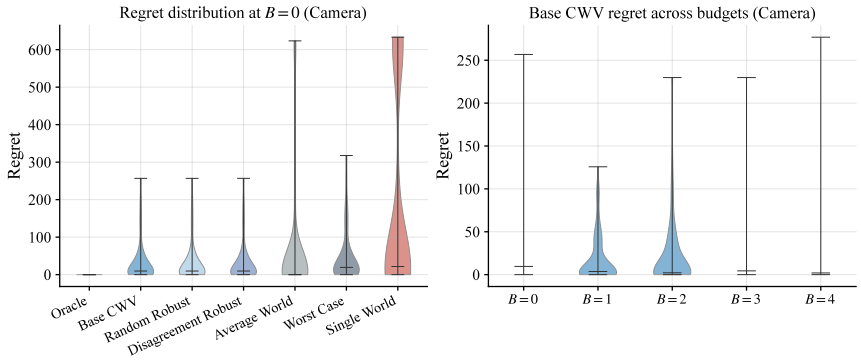

Figure S12: Regret distributions on canonical Berkeley camera ($n{=}50$). Left: distribution at $B{=}0$ across strategies; tails of Single World and Multiverse Vote are dominated by collision penalties. Right: Base CWV's regret distribution across budgets stays narrow with a low median.

E.1 Continuous Planner and Task Tightness

Berkeley Sat3DGen mesh height statistics: min 0 m, max 70 m (after normalization), mean 54.6 m, median 61 m, $p_{90} = 70$ m. RRT* path existence as a function of altitude: at altitudes 60–84 m no feasible route through 20 sampled endpoints exists in 2×10^4 iterations; at altitude 85–90 m all 20 sampled endpoints admit a route with mean clearance ≥ 15 m. The mesh is a solid city block.

E.1.1 Classical Planner Baselines.

To characterize task tightness, we evaluate: grid A* (Hart et al., 1968), PRM (LaValle, 2006), and RRT* (Karaman & Frazzoli, 2011). Below 80 m, all three planners collide in every episode because no obstacle-free path exists in the dense mesh at that altitude and any route intersects a building. At 88 m grid A* succeeds, while PRM and RRT* still exhibit occasional failures because their sampling

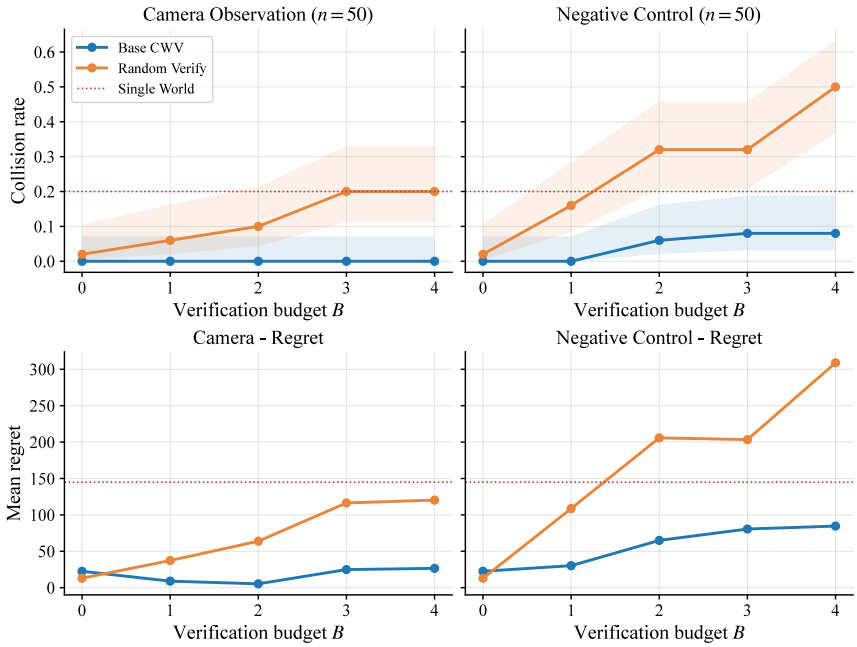

Figure S13: Four-panel statistical summary of the canonical results. Top row: collision rate vs. budget for Base CWV and Random Verify under camera (left) and negctrl (right), with Single World shown as a horizontal reference. Bottom row: mean regret vs. budget for the same strategies. Ribbons are Wilson 95% CI for collision.

Diagnostic	Value
LCB violation rate	0.6% (1/180)
Selected-route collision rate	8.3%
Obs. confidence vs. error	$r=-0.143$, $p=0.056$
Posterior std bin (low): MAE	7.48 m
Posterior std bin (mid): MAE	4.4–5.3 m
Posterior std bin (high): MAE	5.4 m
Post. variance reduction $B=2$	12.2%
Post. variance reduction $B=4$	24.9%

Table S10: Calibration audit on Berkeley negative control ($n=30$). LCB behaves conservatively while posterior standard deviation is inversely calibrated at the low end, isolating the calibration error to the observation posterior.

Setting ($B=2$)	Gated-CWV	Base CWV
Corruption $p=0.75$	**0%** / 31.1 ($\gamma=16$)	3.3% / 51.6
Manhattan (negctrl)	**0%** / 35.1 ($\gamma=16$)	20% / 146.3
Seattle (negctrl)	**0%** / 13.7 ($\gamma=16$)	10% / 76.2
Phoenix (negctrl)	**5%** / 38.5 ($\gamma=16$)	15% / 105.1
Chicago (negctrl)	10% / 76.0 ($\gamma=16$)	**5%** / 58.8
Holdout (negctrl)	**0%** / 17.1 ($\gamma=9$)	10% / 78.7

Table S11: Gated CWV three-way validation.

density does not hug the local minima of building height. Grid A* finds collision-free paths from 88 m onward. Under the tested sampling budgets, PRM and RRT* retain 13–30% collision at 95 m. These results show that altitude dominates safety in the canonical mesh, which lacks an intermediate traversable-but-safety-sensitive regime.

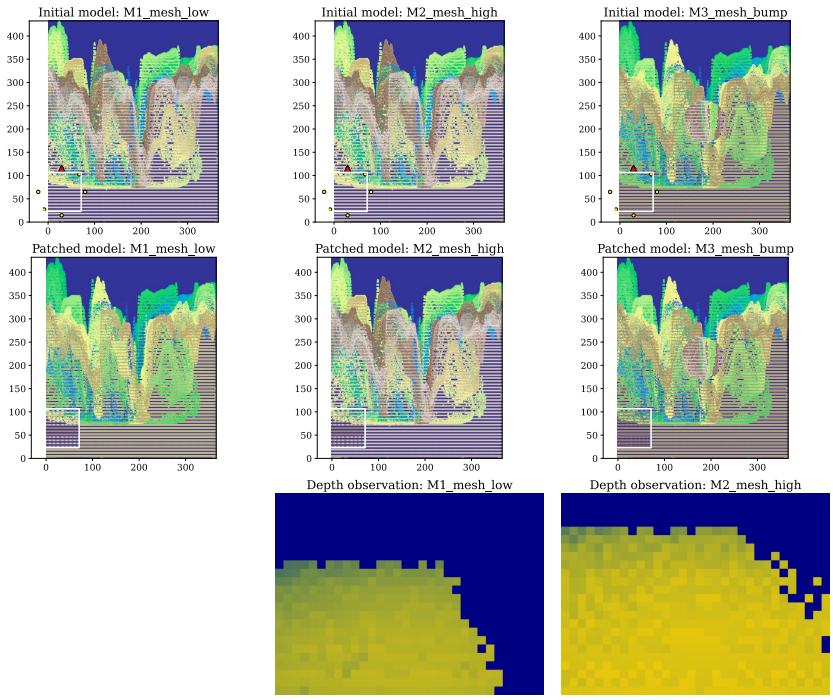

Figure S14: Representative canonical episode. Top: initial route-clearance beliefs under the generated hypotheses. Middle: beliefs after local verification. Bottom: reference clearance observations and the final commitment.

Region	N	Gate fired	Trigger %	% dangerous
Seattle	40	14	35%	14.3%
Chicago	40	8	20%	**87.5%**
Manhattan	40	10	25%	10.0%
Phoenix	40	9	22.5%	22.2%

Table S12: Per-verification diagnostic on the four gated multi-region cities under negctrl at B=2, n=20 per region. "% dangerous" is the fraction of gate-triggered observations that reported a lower clearance than the prior (signed innovation $\eta_r^{\pm} < -\tau$); a symmetric gate discounts these observations even though they carry safety-critical information.

Corruption p	Base CWV		Gated γ=16	
	Coll.%	Regret	Coll.%	Regret
0.00	0.0	5.1	0.0	5.3
0.25	0.0	26.9	0.0	22.1
0.50	3.3	45.9	3.3	45.8
0.75	3.3	51.6	**0.0**	**31.1**
1.00	0.0	44.9	0.0	44.9

Table S13: Corruption sweep on Berkeley camera at B=2, n=30. Gated CWV eliminates the p=0.75 failure and reduces regret at every intermediate corruption level.

E.2 Elevate3D Domain Collapse

Elevate3D on Berkeley: nine axis-aligned orientation checks (transpose, flip x, flip y, rotations at 0/90/180/270 degrees plus their transposes) all yield $|\rho|(\text{Elevate3D}, \text{Sat3DGen}) < 0.15$ at native alignment. Mean height gap on shared (x, y): 41.7 m (Elevate3D 16 m, Sat3DGen 52 m). Route-

Setting	Base	Sym.	Asym.
Berkeley camera, all B	0%	0%	0%
Berkeley NC, B=4	20%	10%	10%
Chicago NC, B=2	10%	5%	**0%**
Chicago NC, B=4	20%	10%	10%
Elevate3D, B=8	80%	80%	80%

Table S14: Safety-asymmetric gate diagnostic. All variants use τ=1 and γ=16 and differ only in the gate condition. Rows are matched within each setting, with n=30 for Berkeley, n=20 for Chicago Loop, and n=5 for Elevate3D.

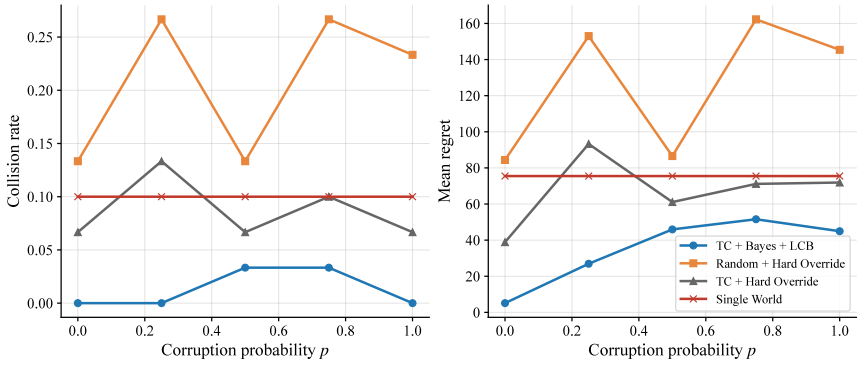

Figure S15: Corruption-mode strategy comparison at B=2, n=30 per corruption level. Naive verification climbs monotonically with p; Base CWV shows a single-episode failure at p=0.5; Gated CWV holds through p=0.75.

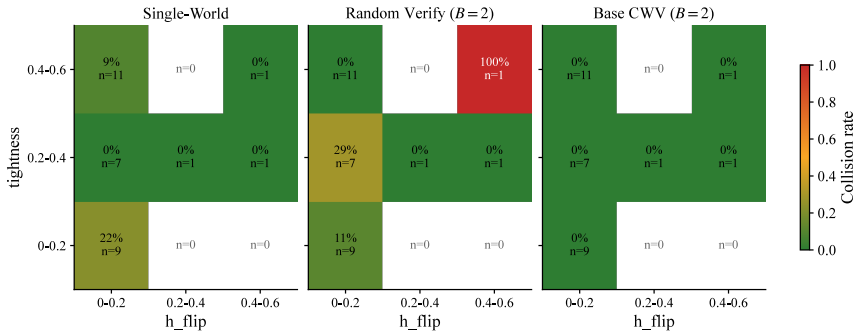

Figure S16: Exploratory phase diagram over scene tightness and reference-flip probability, n=30 per cell. CWV holds at zero across the grid; single-world failures concentrate in the low-tightness low-flip corner.

feasibility split: candidate marks 20 of 20 routes feasible, reference marks 3 of 20. Across five diagnostic episodes, budget swept 0 through 8, collision fraction plateaus at 80% for all strategies.

E.3 Observation Fidelity Sweep

We sweep observation fidelity at three configurations, 16×1 (16-pixel resolution, 1 pose), 24×2 (canonical), and 48×3 (high-fidelity). Under camera observations, the conclusion is unchanged: Base CWV at 0%, naive at 10–40%. Under negative control, Base CWV at B=4 collides at 13.3% at 16×1, 16.7% at 24×2, and 10.5% at 48×3; higher pixel count does not close the residual gap. Table S15 reports the numbers.

Method (B=4)	16×1	24×2	48×3
Camera collision %			
Random Verify	20.0	20.0	20.0
Base CWV	0.0	0.0	0.0
Negctrl collision %			
Random Verify	60.0	64.0	63.2
Base CWV	13.3	16.7	10.5
Gated γ=16	0.0	0.0	0.0

Table S15: Observation fidelity sweep on Berkeley negative-control at B=4 and camera at B=4. Increasing pixel resolution partially reduces but does not eliminate the negctrl gap; the gate closes the residual gap through calibration rather than resolution. n=30 for 16×1 and 24×2; n=19 for 48×3.

E.4 Task-Critical Scoring Sensitivity

The task-critical score uses a bell width ℓ=8 m and three additive offsets of 0.5. We ran a 4×3 factorial sensitivity sweep over $\ell \in \{4, 8, 16, 32\}$ and offsets held equal at $\{0.25, 0.5, 1.0\}$ on Berkeley canonical camera and negative-control at B=2 (n=30 per configuration).

Under camera mode, all 12 configurations produce 0% collision with regret varying by at most 0.1 (5.9 vs 6.0). Under negative control, collision spans 0%–6.7% (0 to 2 out of 30 episodes) across the grid, with the maximum deviation from the default (ℓ=8, offset=0.5) configuration equal to +3.3 pp, within the Wilson noise floor $[0\%, 11.4\%]$ at n=30. Regret varies more visibly under negative control (7.0 to 56.6) but the collision floor is stable. Table S16 reports the negative-control grid.

ℓ \ offset	0.25	0.5	1.0
4	3.3% / 49.4	3.3% / 29.3	0.0% / 24.1
8	3.3% / 33.2	6.7% / 48.6	6.7% / 56.6
16	0.0% / 9.8	0.0% / 7.0	0.0% / 14.0
32	0.0% / 11.3	0.0% / 17.4	3.3% / 30.2

Table S16: Task-critical scoring sensitivity on Berkeley negative control at B=2, n=30 per cell. Format: collision % / mean regret. All 12 configurations lie within the Wilson noise floor $[0\%, 11.4\%]$.

The score behaves as a decision-impact heuristic that CWV's downstream fusion and LCB commit are not sensitive to at the tested resolution.

F Geographic and External Validity

We evaluate alternative geometry sources and candidate constructions beyond the canonical Sat3DGen perturbations.

F.1 Cross-Mesh Details

We use two Sat3DGen meshes generated from different Berkeley satellite inputs, containing 274,564 and 279,186 vertices, respectively. Candidates are constructed as vertex-position mixtures at ratios drawn from $\mathrm{Uniform}[0.2, 0.8]$ per candidate. This produces K=5 candidates each with a distinct position but a shared face topology (both meshes are re-triangulated with matching face count after normalization). No structural error is injected; the only variation across candidates is the mixture ratio.

F.2 Hard Structural Candidates

The main paper briefly describes the default candidate construction, which is detailed here. Each candidate combines five controlled error families: (i) **Large regional systematic bias.** A 2D Gaussian-weighted disk of radius 0.2–0.4 times the map extent centered at a uniformly random

location on the map, with an additive height offset of ± 15 m; used to simulate a systematic elevation error inherited from the input satellite view. (ii) **Missing structures.** 2 to 4 disks of radius 0.05–0.15 times the map extent, each collapsing all vertex heights inside the disk to 10–30% of the original; used to simulate a building or other structure omitted by the generator. (iii) **Hallucinated obstacles.** 2 to 4 disks of radius 0.04–0.10 times the map extent, each raising vertex heights to 60–90% of the reference maximum; used to simulate a hallucinated tall structure. (iv) **Shared systematic bias.** A single additional disk applied identically to all K candidates, biasing the same region in the same direction; used to simulate a common bias across candidates that cross-world disagreement cannot detect. (v) **Per-vertex noise.** Independent Gaussian noise with $\sigma = 3$ m on each vertex.

Table S17 shows the resulting collision rates on this construction. Base CWV attains 0% collision at all budgets. Random Verify with hard-override climbs from 6.7% at B=0 to 20.0% at B=4, reproducing the paradox on structurally perturbed candidates.

Strategy	B=0	B=1	B=2	B=3	B=4
Single World	6.7%	—	—	—	—
Multi Vote	6.7%	—	—	—	—
Random	6.7%	10.0%	13.3%	16.7%	20.0%
InfoGain	0.0%	0.0%	0.0%	0.0%	0.0%
Base CWV	0.0%	0.0%	0.0%	0.0%	0.0%

Table S17: Hard structurally perturbed candidates on Berkeley, camera mode, n=30. — indicates budget-independent strategies. Base CWV holds at zero across budgets; naive verification worsens.

F.3 Cross-Source OSM vs. Sat3DGen

The OSM procedural extrusion uses OpenStreetMap building footprints only, extruded to a default height derived from tag data or a per-region default. Trees, terrain slopes, and non-building surface features are absent. On Berkeley the OSM mesh has 2,187 vertices; Sat3DGen has 274,564. We rasterize both into a shared coordinate frame at the same $[0, 70]$ m height scale. Mean height on shared (x, y): OSM 27.2 m, Sat3DGen 52.5 m; systematic gap -25.4 m. Per-candidate perturbations add $\sigma = 3$ m Gaussian vertex noise plus 1–3 localized structural errors (missing or hallucinated).

F.4 LiDAR Terrain Evaluation

USGS 3DEP source: `CA_AlamedaCounty_2021_B21` product; source URL structure follows the pattern documented at (U.S. Geological Survey, 2023). Coordinate system: EPSG:26910 (UTM Zone 10N, NAD83, NAVD88). Crop bounds: easting $[562840, 563160]$, northing $[4191340, 4191660]$ (320 m×320 m). Downsampling: 320×320 native (1 m) → 80×80 (4 m) via mean pooling. Mesh conversion: rectangular grid → triangulated surface (6,400 vertices, 12,482 faces). Conversion to the pipeline's normalized frame: height scale 70 m preserved. The downtown Berkeley tile has a total vertical range of $\approx$ 12 m, which places every candidate route in a permissive-margin regime; no strategy at any budget produces a collision on this LiDAR terrain. High-budget degradation therefore manifests only in the regret tail rather than in collision counts: Random Verify's worst-episode regret grows non-monotonically with budget (14.88, 11.32, 19.33, 11.32, 28.32 at B=0, 1, 2, 3, 4), peaking sharply at B=4, while Base CWV's mean regret holds in the 2.35–4.39 range across budgets (Table S18). This regret-tail signature is distinct from the collision-based Verification Paradox observed on tighter geometries.

G Representation Extensions

We extend the analysis to implicit video rollouts and a strongly conditioned density-field generator.

G.1 DreamX Representation Extension

DreamX uses the Wan2.2-TI2V-5B base model with the DreamX-World-5B-Cam adapter, generating 81-frame rollouts at 704×1280 resolution and 16 FPS with 50 diffusion steps.

Strategy	B=0	B=1	B=2	B=3	B=4
Base CWV	3.23	2.35	2.35	2.35	4.39
Random Verify	0.93	0.93	1.33	0.74	3.63
Disagreement	0.93	0.59	0.09	0.09	5.30
Single World	1.00	1.00	1.00	1.00	1.66
Rand. max regret	14.88	11.32	19.33	11.32	**28.32**

Table S18: USGS LiDAR bare-earth Berkeley, n=30. All strategies at all budgets record 0% collision on this flat terrain; the paradox manifests only in the regret tail. Random Verify's worst-episode regret at B=4 (28.32) shows a naive-verification failure mode that the mean does not capture. Base CWV's regret stays bounded across budgets.

Cross-World Consistency Frontier metric: composite $= 0.6 \cdot (1 - \text{SSIM}) + 0.25 \cdot \text{edge_div} + 0.15 \cdot \text{pixel_L}_1$, computed pairwise across seeds at each frame. Frontier: first frame at which composite ≥ 0.35 and remains ≥ 0.315 for at least 3 subsequent frames. Frontier time by scene: coast 1.25 s (frame 20), pagoda 1.88 s (frame 30), Minecraft 3.63 s (frame 58). Cross-seed disagreement growth from t=0 to peak: 87.6× (coast), 61.5× (pagoda), 81.4× (Minecraft).

CWV pipeline on DreamX: 7 commitment depths (frames 5, 17, 30, 42, 54, 66, 76), K=4 seeds, R=7 routes, $B \in \{0, 1, 2, 3, 4\}$, n=50 episodes with observation noise. Camera-mode collision: Base CWV 0% at all budgets; Single/Average/Worst 100%; Random Verify 94–100%. Negctrl-mode collision: Base CWV 34%–84% (increasing in budget); Random 98–100%.

Scene	t=0	Frontier (s)	Growth (×)
Coastal cliff	0.0051	1.25	87.6
Pagoda hillside	0.0052	1.88	61.5
Minecraft cliff	0.0048	3.63	81.4

Table S19: DreamX cross-world consistency frontier.

G.2 Sat2Density++ Positive Control

Sat2Density++ generation setup: satellite image (Chicago from VIGOR dataset), driving trajectory, sky panorama; NeRF density field via marching cubes to a colored mesh (Lorensen & Cline, 1987; Schroeder et al., 2006). We generate four Sat2Density++ meshes from the same conditioned input.

Metric	Value
Seeds	$\{0, 1, 2, 3\}$
Vertex count per mesh	141,307–143,679
Mean Z per mesh	19.4–19.5 m
Cross-seed clearance std mean	0.24 m
Cross-seed clearance std max	0.31 m
Range across seeds	$\leq$ 0.7 m

Table S20: Sat2Density++ positive control on Chicago. Cross-seed clearance is negligible relative to the safety margin (16 m); the uncertainty signal does not falsely trigger under a strongly conditioned generator.

H Generated-Future Actionability in a Multimodal Creative Agent

The main experiments study generated worlds as hypotheses about physical geometry. We extend the same actionability question to a multimodal creative agent that plans over generated visual futures while maintaining an editable artifact state. Here the hypothesis is an action-conditioned preview rather than a 3D scene. Separating preview generation, production execution, and future editing lets us distinguish present quality from future controllability and prediction–execution calibration.

The agent maintains an executable artifact state and can patch, branch, verify, commit, and continue editing. For currently feasible states, Prospective Editability Value (PEV) augments immediate utility with expected downstream edit utility: whether a future request can be executed locally, whether

protected content is preserved, and whether generative fallback is required. Prediction–Execution Calibration (PEC) separately measures whether a generated preview predicts the consequence of production execution.

H.1 Prospective Editability under Identical Current Appearance

We first remove current visual quality as a confound. In a same-pixels semantic-flattening intervention, each accepted artifact is represented either as an executable state with live semantic entities or as an opaque raster containing exactly the same visible pixels. Current appearance and current utility are therefore identical; only the future action surface differs. Across four diagnostic tasks, prospective replay selects the executable state in all four cases. Future local-edit success is 1.0 for the executable state and 0.0 for the flattened state, while expected generative fallback is 0.0 versus 1.0.

A held-out next edit changes a price from `$39` to `$35` and a CTA from `SHOP` to `BUY NOW`, while preserving the product and brand. The executable state applies the revision locally with zero image-model calls, 0.046 s latency, and protected-product exact-pixel rate 1.0. The flattened raster requires one generative image edit, takes 12.387 s, preserves only 0.340 of protected pixels exactly, and duplicates the requested CTA.

	Executable state	Flattened raster
Current pixels	identical	identical
Current utility	identical	identical
Future local-edit success	**1.0**	0.0
Expected image fallback	**0.0**	1.0
Selected by prospective value	**4/4**	0/4

Table S21: Same-pixels intervention: present appearance is fixed; only the future action surface changes.

H.2 When Prospective Value Changes the Current Decision

A four-task pool with 15 valid candidates yields five strict same-task pairwise current-versus-PEV reversals at $\lambda{=}0.25$. Its feasibility-gated global selector changes top-1 on $0/4$ tasks because each task already contains a high-current-quality executable state. A harder pre-registered five-task suite therefore tests the same question under stronger current-quality competition.

The harder suite contains 29 valid real candidates and exactly one strict current-feasible global top-1 disagreement. On the app-launch task, current-only selection prefers a raster state with immediate utility 0.953111 and prospective value 0.200000, while PEV selects an executable state with lower immediate utility 0.922500 but prospective value 1.000000. The current-score sacrifice is 0.030611 and the prospective gain is 0.8.

The pair is frozen before revealing the pre-authored next request. The executable state completes the edit with zero image-model calls, 0.225 s latency, exact copy integrity, and exact protected/non-target pixels. The current-only raster winner requires one image-model call and 13.386 s, fails copy integrity, and changes nearly all protected and non-target pixels exactly.

	Current-only top-1	PEV top-1
Representation	raster	executable
Immediate utility	**0.953111**	0.922500
Prospective value	0.200000	**1.000000**
Future image-model calls	1	**0**
Future latency (s)	13.386	**0.225**
Exact copy integrity	fail	**pass**
Protected exact pixels	0.008872	**1.000000**
Non-target exact pixels	0.033332	**1.000000**

Table S22: Pre-registered global top-1 disagreement followed by the frozen held-out edit. One of five task families exhibits the selector disagreement.

The result separates present-quality ranking from future-actionability ranking under a fixed feasibility rule: the candidate preferred now need not be the state that best supports the next interaction.



Figure S17: **Present quality versus future editability.** Top: the higher-current raster candidate (left) and the PEV-favored executable candidate (right). Bottom: the frozen held-out edit. Raster regeneration changes broad non-target content and leaves the obsolete CTA, while the executable state applies the requested price/CTA change locally.

H.3 Prediction–Execution Calibration

Generated previews are useful for planning only if they predict the consequence of the real executor. We therefore hold semantic actions and production outcomes fixed and vary preview fidelity. An earlier scalar-only check assigned the same quality score (0.9575) to preview and production even when pairwise inspection found a production-only artifact, motivating consequence-level PEC rather than independent scalar scoring.

On the headphones task, 384-, 512-, and 768-pixel previews all preserve the production ranking with zero measured production regret. On the event task, the calibration signal changes with fidelity: 384 preserves ranking but has selected Prediction–Execution Gap (PEG) 0.052536 and consequence agreement 0.925; 512 flips an almost tied production ranking with regret 8×10^{-6} and artifact mismatch 0.05; 768 restores the ranking with zero measured regret, selected PEG 7.8×10^{-5}, consequence agreement 1.0, and zero artifact mismatch.

The sweep treats preview fidelity as a control variable: a cheap prediction is sufficient on the stable family, while the less stable event task benefits from stronger prediction before commitment.

Task	Preview	Rank	Prod. regret	Sel. PEG	Conseq. agr.	Artifact
Headphones	384	1.0	0	0	1.000	0
Headphones	512	1.0	0	0	1.000	0
Headphones	768	1.0	0	0	1.000	0
Event	384	1.0	0	0.052536	0.925	0
Event	512	0.0	8×10^{-6}	0.037507	0.975	0.05
Event	768	1.0	0	0.000078	1.000	0

Table S23: Fixed-action prediction–execution fidelity diagnostic. Semantic actions and production outcomes are held fixed while preview fidelity varies.

H.4 Execution Experience and Adaptive Test-Time Compute

Execution history also changes how much generated-future compute is useful. On a stable headphones family, experience reduces the rollout budget from $K{=}3$ to $K{=}2$; realized image calls fall from 3 to 2 and the selected score remains 0.9575. On a failure-prone coffee family, experience increases the budget from $K{=}2$ to $K{=}3$, raising the selected score from 0.9225 to 0.9575.

Thus the adaptive policy matches fixed-$K{=}3$ quality with one fewer image call and improves over fixed-$K{=}2$ by 0.0175 mean selected score for one additional call. A provider-backed episode exercises the three signals jointly: retrieved failure history changes K from 2 to 3, consequence risk selects a 768-pixel preview, future-edit importance sets the PEV weight to 0.40, and pairwise PEC is enabled. The realized trajectory follows these planned settings and commits a result with score 0.9575, exact protected content, consequence agreement 1.0, and PEG 0.0.

Policy	Calls	Mean score	Exact text
Fixed $K{=}2$	4	0.9400	1.0
Fixed $K{=}3$	6	**0.9575**	1.0
Experience-adaptive	**5**	**0.9575**	1.0

Table S24: Adaptive compute over two held-out task families.

Connection to generated-world actionability. The physical and creative settings share the same decision pattern: generated hypotheses predict task-relevant consequences, external evidence calibrates those predictions, and commitment acts on the resulting uncertainty. In the creative agent, prospective editability adds a temporal dimension by valuing the future action surface of the committed state. Together, the results extend generated-world actionability from spatial hypotheses to generated action consequences.