IMAGE-DEFECT-GUIDED CONFIGURATION OF FROZEN VIDEO GENERATORS FOR AUTONOMOUS-DRIVING EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Driving-video generators provide synthetic observations for offline evaluation of autonomous-driving systems, but visually plausible outputs can alter perception, tracking, and planning metrics relative to condition-aligned real observations. Inference-time controls offer a way to reduce these discrepancies without retraining, yet evaluating every configuration with the full driving stack is costly. We present a reference-based configuration workflow that uses image-defect measurements to prioritize downstream evaluation while keeping both the generator and evaluators frozen. Single-factor screening identifies controls that alter task-associated defects, and a lightweight surrogate ranks candidates within the retained search space. Each search round sends only the best-ranked candidate for full downstream evaluation, while a separate multi-task check assesses nominated configurations on held-out scenes. Experiments with the WorldDreamer renderer and a Cosmos-based pipeline on 26 InterHub scenes show that reconstruction and temporal-consistency defects are associated with perception and planning gaps in both pipelines after accounting for scene and generator effects. Held-out evaluation identifies one configuration per pipeline that reduces the composite gap while satisfying task-specific tolerance checks under the frame-level decision rule. The relative reductions are 31.24% and 21.02% for WorldDreamer and Cosmos, respectively, corresponding to absolute reductions of 0.1415 and 0.0952 on the fixed percentile-normalized scale. The WorldDreamer configuration primarily narrows the tracking gap, while the Cosmos configuration narrows both perception and tracking gaps. In a matched single-run comparison, guided search reaches a lower task gap using 41% of the GPU time recorded for random search. These results support image-defect-guided configuration of frozen generators for offline driving evaluation, with image measurements prioritizing candidates and observed task-metric gaps determining which configurations to retain.

1 INTRODUCTION

Controllable driving world models synthesize videos from maps, trajectories, agent layouts, and camera conditions (Guan et al., 2025; Wang et al., 2024a;b; Yang et al., 2025; Gu, 2025). These generated observations provide inputs for evaluating and training autonomous-driving systems (Yang et al., 2025; Yan et al., 2025). For offline evaluation, their usefulness depends on whether a fixed driving stack produces comparable task metrics from generated and real observations of the same scenario. A plausible video may nevertheless omit an actor, disrupt an identity, or distort geometry and motion cues. Such defects can change perception, tracking, and planning metrics and complicate interpretation of the evaluation.

We focus on settings in which the generator and downstream evaluators remain frozen. The available decisions concern exposed inference-time controls, including guidance scale, temporal memory, and denoising range. Selecting these controls requires task feedback, but evaluating every candidate with the full driving stack is costly. We therefore ask how image-level measurements can prioritize configurations that improve agreement with real-reference task metrics while reducing the number of downstream evaluations.

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.

We define task fidelity as agreement in scalar task metrics between generated observations and condition-aligned real references under a fixed evaluator stack (Section 3.1). This definition does not imply identical detections, track identities, or planned trajectories, nor does it establish equivalence in closed-loop driving performance. Paired real observations are required during configuration selection, making our setting a form of reference-based offline calibration. Research on visual similarity, downstream evaluation, and task-guided simulation provides foundations for this problem (Wang et al., 2004; Zhang et al., 2018; Chen et al., 2021; Lindström et al., 2024; Wang et al., 2025; Mishra et al., 2022). We build on these directions to study which exposed controls merit search and how observed task gaps should determine the final configuration decision.

This problem requires three distinctions between the evidence used at different stages. First, real and generated observations must be aligned by scenario, trajectory, view, and time to separate scene difficulty from generation-related discrepancies. Second, perception, tracking, and open-loop planning can respond differently to generated inputs, so the task-gap vector matters alongside its aggregate. Third, an image defect associated with task fidelity need not respond to an available control, and a control-induced change need not improve every task. We therefore assess diagnostic associations, local control sensitivity, and observed task-gap changes on held-out scenes as separate forms of evidence.

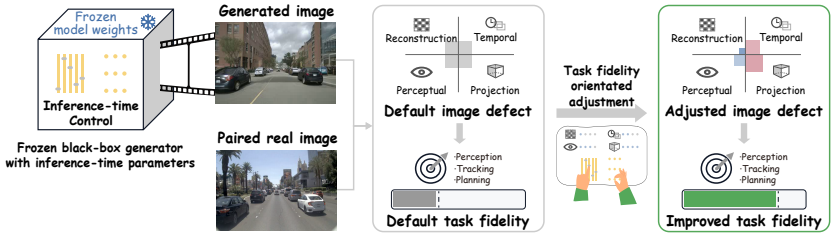

Figure 1: Motivation. A frozen driving-video generator supplies observations to a frozen autonomy stack, and configuration changes are restricted to its inference-time controls. Image defects measured on aligned pairs of real and generated observations prioritize configurations for evaluation. Observed perception, tracking, and planning gaps determine whether a configuration meets the stated decision rule.

Figure 1 summarizes how image defects provide evidence for candidate prioritization while observed downstream task gaps determine the final decision. We study this workflow for auditing, screening, and confirming configurations with WorldDreamer and a Cosmos-based pipeline on aligned InterHub scenarios (Yang et al., 2025; Gu, 2025; Jiang et al., 2025). RQ1 asks which image and video defects are associated with downstream task-metric gaps. RQ2 asks whether these measurements preserve task-gap ordering in the held-out diagnostic sample and provide evidence for configuration prioritization. RQ3 asks which nominated configurations meet the stated multi-task decision rule on held-out scenes.

These experiments yield the following contributions to configuring frozen driving-video generators.

- A condition-aligned audit identifies reconstruction and temporal-consistency defects associated with perception and planning gaps in both pipelines after accounting for scene and generator effects. No tracking association meets the stated promotion criteria in this audit.
- A configuration workflow combines defect-based candidate prioritization with a separate multi-task decision rule. It uses surrogate modeling and optimization to address configuration of frozen generators, while separating predicted priorities from observed task outcomes.
- Experiments identify one configuration per pipeline that meets the stated frame-level rule, with composite-gap reductions of 31.24% and 21.02% relative to the respective defaults. A single-run comparison with random search provides descriptive evidence on search efficiency.

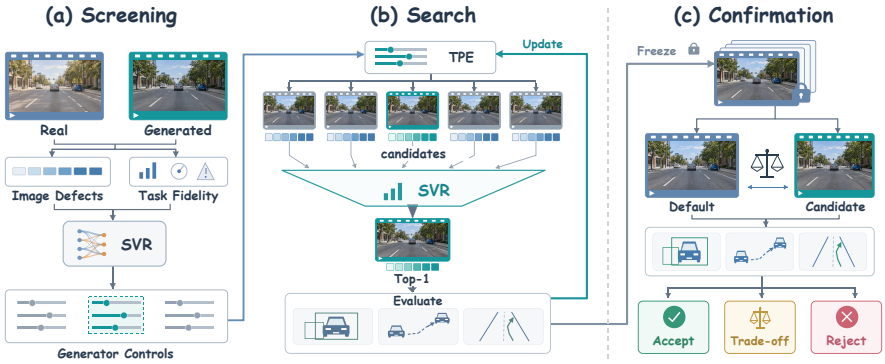

Figure 2: Workflow for auditing, screening, and confirming configurations. Paired defect vectors and composite task gaps train the ranking surrogate, while residualized associations identify guardrail channels. Single-factor screening selects controls, and local search evaluates the best-ranked candidate from each batch. Observed task gaps update the search center, and held-out evaluation determines the final decision.

2 Related Work

Controllable driving world models and sensor simulation. Conditional video generators synthesize driving observations from maps, layouts, trajectories, actions, or reference frames, with increasingly structured, multi-view, and temporally consistent generation mechanisms (Wang et al., 2024a;b; Wen et al., 2024; Lu et al., 2024; Gao et al., 2024b; Zhao et al., 2025). A complementary line builds neural sensor simulators from reconstructed scenes or couples learned generation with interactive simulation for closed-loop evaluation (Yang et al., 2023; Tonderski et al., 2024; Yang et al., 2025; Yan et al., 2025). Our work proposes neither a new generator nor a new simulator; it audits trained generators as black boxes and asks how their exposed controls can be screened using downstream evidence. Throughout this paper, WorldDreamer denotes the diffusion-based World Dreamer renderer in DriveArena (Yang et al., 2025), and Cosmos denotes our multi-view pipeline built on the Cosmos world foundation model platform (Gu, 2025; NVIDIA, 2025).

Task-oriented fidelity and real-to-simulation utility. Visual metrics such as the structural similarity index measure (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) quantify structural or perceptual similarity efficiently, and distribution-level scores such as FID and FVD summarize realism over a corpus (Wang et al., 2004; Zhang et al., 2018; Heusel et al., 2017; Unterthiner et al., 2018), but none of them tests whether an autonomy stack preserves its behavior on a given scenario. Generated driving data are therefore often judged by downstream utility, for example by training or evaluating detectors on generated images (Wang et al., 2024a; Wen et al., 2024; Gao et al., 2024a), and later studies connect synthetic sensor data and neural-rendering quality with downstream task performance (Chen et al., 2021; Lindström et al., 2024). Most closely related, Drive&Gen compares matched real and generated planner responses and varies operational conditions to diagnose planner domain gaps (Wang et al., 2025). Its interventions change scene conditions and target an end-to-end planner. By contrast, we keep scenario conditions aligned while screening inference-time controls of a fixed generator, and confirm candidates with a vector spanning perception, tracking, and open-loop planner response. Our focus is the generator-configuration decision that follows task-aware evaluation, not a universal benchmark score.

Interactive driving evaluation and task-guided simulation. In autonomous-driving simulation, a closed loop conventionally means that agent actions change simulator state and hence subsequent observations, as instantiated at different levels of sensor realism by NeuroNCAP, Bench2Drive, nuPlan, DriveArena, and DrivingSphere (Ljungbergh et al., 2024; Jia et al., 2024; Karnchanachari et al., 2024; Yang et al., 2025; Yan et al., 2025). Task2Sim instead learns simulation parameters that generate useful data for a target recognition task (Mishra et al., 2022). Our setting neither places the driving agent in an interactive simulator nor retrains the generator; it forms an audit–screen–confirm loop around paired replay, in which sparse control sweeps test defect hypotheses locally and separate downstream evaluation returns the decision.

Actionable black-box configuration. Inference-time settings of diffusion generators, such as the classifier-free guidance scale, are known to trade sample fidelity against diversity (Ho & Salimans, 2022); we treat such settings as black-box knobs whose downstream effect is measured rather than modeled. Predictive explanation and causal explanation answer different questions (Shmueli, 2010), and permutation-style importance can become unstable under correlated predictors (Breiman, 2001; Strobl et al., 2008; Hooker et al., 2021), so we read defect-feature importance only as an association with composite degradation under the fixed evaluation stack. Our sparse control sweeps relate to elementary-effects factor screening (Morris, 1991; Campolongo et al., 2007), and the surrogate-guided search parallels surrogate-assisted and multi-objective optimization (Booker et al., 1999; Deb et al., 2002). We measure local response sensitivity over prespecified control ranges, without claiming global sensitivity, causal attribution, convergence to an optimum, or Pareto optimality.

3 METHOD

Figure 2 separates screening (a), configuration search (b), and downstream confirmation (c). Image defects prioritize candidates; only observed downstream gaps update the retained configuration and search center.

3.1 CONDITION-ALIGNED PAIRED AUDIT

Let m index a frozen generator with controls θ and default θ_0. Condition c_i specifies trajectory and road inputs for observation i, aligned by scene, timestamp, and camera view. WorldDreamer additionally uses a fixed built-in style reference. Cosmos generates a front view from text and HD-map control video, then uses the generated front view with multi-view controls to synthesize the remaining views. Paired real frames serve as evaluation references rather than scenario-specific generator inputs. We compare real x_i with generated $\widehat{x}_i^{(m)}(\theta) = g_\theta^{(m)}(c_i)$, fixing weights, input conditions, evaluators, seeds, and alignment across configurations.

Let $\mathcal{K} = \{\mathrm{per}, \mathrm{trk}, \mathrm{plan}\}$ denote the perception, tracking, and planning tasks. We operationalize task fidelity through scalar task-metric gaps between frozen downstream evaluator outputs on generated observations and their condition-aligned real references:

$$G_{i,k}(\theta) = \left| q_k(x_i) - q_k(\widehat{x}_i(\theta)) \right|, \quad k \in \mathcal{K}. \tag{1}$$

Here q_k uses F1, window-level Higher Order Tracking Accuracy (HOTA), and trajectory L_2 error for perception, tracking, and planning. Zero denotes agreement; larger gaps denote lower fidelity. The direction-aligned defect vector $\mathbf{z}_i = [\mathrm{L1}, -\mathrm{PSNR}, \mathrm{LPIPS}, \mathrm{ProjErr}, \mathrm{LPIPS\text{-}t}, \mathrm{FlowErr}]^\top$ spans reconstruction, perceptual, projection, and temporal fidelity. L1 and PSNR measure reference-based pixel discrepancies that can reflect both appearance and structural differences. Because the conditioning does not uniquely determine appearance, these quantities serve as empirical proxies rather than pure measures of structural fidelity. Each task gap is normalized to $[0, 1]$ by the training-fold 5th and 95th percentiles of that task's gap and clipped, giving $\widetilde{G}_{i,k}$; their equal-weight mean $a_i = \frac{1}{3}\sum_k \widetilde{G}_{i,k}$, averaged over valid observations as $A(\theta)$, is used only for ranking and aggregate comparison. We call this the *percentile-normalized composite scale*; composite gains are differences of means on it, not standard-deviation units.

3.2 RESIDUALIZED DEFECT–TASK ASSOCIATION

For each defect j and task k, we rank both variables and remove additive generator and scene effects:

$$\rho_{j,k} = \mathrm{Spearman}\left(\mathrm{resid}[\mathrm{rank}(z_{i,j})], \mathrm{resid}[\mathrm{rank}(G_{i,k})]\right). \tag{2}$$

Spearman is applied to residuals of ranked variables, distinct from Pearson correlation of rank residuals. Benjamini–Hochberg correction controls false discovery. A defect enters $\mathcal{J}_{\mathrm{guard}}$ only when its association is positive and significant and remains positive within both generators.

3.3 SURROGATE-GUIDED CONFIGURATION SCREENING

Stage 1: single-factor screening. For control p and tested value $v \in \mathcal{V}_p$, only $p = v$ changes; other controls stay at θ_0, with matched scenes and seeds. The screening procedure is detailed in

Appendix B. The normalized default-relative defect response is

$$\Delta z_{p,j}(v) = \frac{1}{|\mathcal{I}||\mathcal{R}|} \sum_{i\in\mathcal{I}} \sum_{s\in\mathcal{R}} \left[\widetilde{z}_{i,s,j}(p{=}v) - \widetilde{z}_{i,s,j}(\theta_0)\right], \tag{3}$$

where $\mathcal{I}$ is the matched sweep-observation set, $\mathcal{R}$ is the repeated-seed set, and $\widetilde{z}$ uses frozen reference-cohort normalization.

An RBF-kernel support vector regressor (RBF-SVR), fitted with scene-grouped validation, maps absolute defect vectors to composite gaps. The function f includes fitted imputation and standardization, frozen at inference. The configuration prediction is

$$\widehat{a}_p(v) = \frac{1}{|\mathcal{I}||\mathcal{R}|} \sum_{i\in\mathcal{I}} \sum_{s\in\mathcal{R}} f(\mathbf{z}_{i,s}(p{=}v)).$$

The default prediction uses the same observations, seeds, and transform. We compare these predictions; $\Delta\mathbf{z}_p(v)$ describes mobility and is not an SVR input. Scene-held-out ordering supports defects as a low-cost proxy for configuration prioritization, not a replacement for downstream evaluation.

We score response magnitude and scene–seed stability jointly:

$$S_p = \max_{v\in\mathcal{V}_p} \mathcal{N}_{\mathrm{ref}}\left(|\widehat{a}_p(v) - \widehat{a}(\theta_0)|\right) \mathrm{Stab}_p(v), \tag{4}$$

Here $\mathcal{N}_{\mathrm{ref}}$ uses frozen reference statistics and $\mathrm{Stab}_p(v) \in [0,1]$ measures repeat agreement across scenes and seeds. A control is retained when $S_p \geq S_{\min}$ and at least one promoted channel $j \in \mathcal{J}_{\mathrm{guard}}$ shows reproducible movement $\Delta z_{p,j}(v)$ in Equation 3. The stability definition, reference normalization, and retention threshold are specified in Appendix B. Retained axes define $\Omega_m = \prod_{p\in\mathcal{C}_{\mathrm{ret},m}} [\ell_p, u_p]$.

Stage 2: local candidate search. On the fixed search cohort $\mathcal{I}_{\mathrm{search}}$, let $\widehat{a}(\theta) = |\mathcal{I}_{\mathrm{search}}|^{-1} \sum_i f(\mathbf{z}_i(\theta))$, using the frozen screening-fitted transform. Candidates are ranked by

$$R(\theta) = \widehat{a}(\theta) + \lambda \sum_{j\in\mathcal{J}_{\mathrm{guard}}} \max\{0, d_j(\theta) - \tau_j\}. \tag{5}$$

Here d_j is normalized default-relative defect change, τ_j its tolerance, and λ the penalty weight. A local tree-structured Parzen estimator (TPE) proposes five configurations around $\theta_m^{(r)}$ (Bergstra et al., 2011). All are ranked by R, but only the lowest-R candidate is evaluated on all search-cohort frames. An observed composite-gap improvement updates both the retained configuration and $\theta_m^{(r+1)}$; otherwise both remain unchanged. Search runs for five WorldDreamer and 14 Cosmos rounds, without implying convergence. Final decision labels do not feed back into ranking.

3.4 Multi-Task Confirmation

Candidate identities and order are frozen before comparing each candidate with its default under matched scenes, seeds, and evaluators. For valid paired frames $\mathcal{I}_k$, gains are

$$\begin{aligned} \Gamma_k(\theta) &= \frac{1}{|\mathcal{I}_k|} \sum_{i\in\mathcal{I}_k} \left[G_{i,k}(\theta_0) - G_{i,k}(\theta)\right], \\ \Gamma_0(\theta) &= A(\theta_0) - A(\theta). \end{aligned} \tag{6}$$

Positive values indicate smaller gaps. Let L_0, L_k be lower endpoints of nominal paired 95% frame-bootstrap intervals and μ_k task-specific non-inferiority margins from identical-input repeats.

The final decision is

$$D(\theta) = \begin{cases} \text{accept}, & \substack{L_0(\theta)>0,\\ L_k(\theta)\geq -\mu_k,\quad \forall k\in\mathcal{K},} \\ \text{trade-off}, & \substack{\exists k\colon L_k(\theta)>0,\\ \exists \ell\colon L_\ell(\theta)<-\mu_\ell,} \\ \text{reject}, & \text{otherwise.} \end{cases} \tag{7}$$

Acceptance requires a positive nominal aggregate-gain lower bound and component-wise bounds within the stated tolerances. A trade-off denotes a positive nominal gain bound in at least one

component with the tolerance condition unmet in another, not an established deterioration. Reject denotes that neither condition holds. These are operational decisions under a frame-level rule, not statistically established scenario-level improvement or non-inferiority. No multiplicity correction is applied across the 19 evaluated candidates.

4 Experimental Setup

Generators and aligned scenarios. We evaluate the DriveArena WorldDreamer renderer and our Cosmos-based pipeline on 26 interaction-rich scenes selected from InterHub (Yang et al., 2025; Gu, 2025; Jiang et al., 2025). Real references are nuScenes images corresponding to the selected InterHub trajectories (Caesar et al., 2020). Fixed-interval sampling yields 864 aligned image-defect observations (432 per generator), aligned by scene, timestamp, and camera view. Scenes define data partitions, while individual observations provide evaluation samples. Only native inference-time controls are varied; architecture and weights remain fixed.

Fixed downstream evaluators. The same evaluator stack is applied unchanged to real references, default generations, and candidates. Perception uses BEVFormer (Li et al., 2022), tracking uses YOLOv8n detections with ByteTrack association (Zhang et al., 2022), and open-loop planning uses VAD (Jiang et al., 2023). Their F1, HOTA, and planning trajectory L_2 outputs define the absolute task-fidelity gaps in Equation 1. These endpoints cover perception performance, detection and temporal association quality, and planner output deviation. On real references from three scenes, F1 spans 0.000–0.788 and planner L_2 spans 9.56–32.93 m. We interpret paired real–generated differences under the same frozen evaluator rather than these absolute levels as driving capability. Real–real repeats establish evaluator-repeatability margins for confirmation. Annotation sources, planning references, checkpoints, thresholds, and frozen configurations are specified in Appendix B.

Diagnostic modeling and data separation. The 26 scenes are partitioned into 12 screening, six search, and eight confirmation scenes, with all frames, camera views, seeds, and generator versions of each scene assigned to the same cohort. The six defect channels are fitted with an RBF-SVR under five-fold scene-grouped validation within the screening cohort; imputation, standardization, task-gap normalization, and hyperparameter selection use training-fold statistics. Search uses the six search groups, and candidates are frozen before evaluation on the eight held-out confirmation groups. The reported statistics follow this 12/6/8 scene-disjoint protocol; confirmation intervals use paired frame bootstrap.

Control screening and confirmation. Single-factor sweeps and local TPE search follow Section 3.3; matched-frame confirmation follows Section 3.4. Appendix A lists native controls and search ranges (Table A1) and their defaults.

5 Results

5.1 Frame-Level Paired Audit

Condition-aligned auditing confirms that both generators exhibit measurable image defects and downstream task-fidelity gaps, with non-zero variation in all six defect channels and all three task gaps.

Table 1: Frame-level paired audit. Defect columns report mean discrepancies, and task columns report non-negative absolute gaps in evaluator units. Smaller is better, except that higher PSNR is better.

	Image-defect measurements						Absolute task-fidelity gaps		
Generator	L1	PSNR	LPIPS	Proj.	Flow	LPIPS-t	F1 gap	HOTA gap	Plan. gap (m)
WorldDreamer	0.1966	11.6525	0.6018	0.9680	1.1205	0.1028	0.2858	0.2009	2.5341
Cosmos	0.2170	11.3198	0.6933	0.9893	1.8075	0.0994	0.3035	0.1661	1.6539

Cosmos has larger mean L1, LPIPS, projection, and flow discrepancies and a larger F1 gap, whereas WorldDreamer has larger HOTA and planning gaps. Larger image defects thus need not imply

larger gaps on every task, which motivates retaining the task-gap vector rather than treating any image metric as a task-fidelity score.

Evaluator sanity checks. Repeated identical-input runs on three real scenes set non-inferiority margins of 0.0452 F1, 0.0000 HOTA, and 0.0031 m from the maximum 95th-percentile absolute frame differences. The estimated HOTA margin is zero, so the rule requires its nominal gain lower bound to be nonnegative. These margins are component-wise decision tolerances, not driving thresholds; they define whether a candidate passes the stated repeatability-based rule.

5.2 Defect–Task Associations Consistent Across Both Pipelines

Reconstruction and temporal-consistency defects retain significant, non-reversing associations with task-fidelity loss in both generators under the promotion criteria in Section 3.2 (Appendix C, Table A4).

L1, direction-aligned PSNR, and LPIPS-t show the strongest planning associations ($\rho = 0.502$–0.548, $q < 10^{-7}$) and significant perception associations ($\rho = 0.320$–0.366); flow error provides weaker planning evidence ($\rho = 0.281$, $q = 5.9 \times 10^{-3}$). LPIPS and projection error are not promoted, and no defect shows a stable association with the tracking gap.

The association analysis does not establish a stable defect-based predictor of tracking gaps. Tracking outcomes therefore require direct downstream evaluation in our workflow. Image defects are frame-level aggregates, whereas HOTA depends on detection and temporal identity association, so localized visual continuity can change tracking without a pooled defect correlation.

5.3 Inference-Time Control Sensitivity

Frozen-generator controls alter task-associated defect channels. Default-relative responses $\Delta \mathbf{z}_p(v)$ show that WorldDreamer guidance scale and temporal memory, and Cosmos front-view partial denoising, produce their largest displayed responses in L1 and flow error (Figures A3 and A2).

5.4 Configuration Screening and Guardrail-Aware Ranking

Defect-association surrogate. On scene-held-out data, the surrogate preserves the ordering of composite task-fidelity gaps (Spearman $\rho = 0.8338$, $R^2 = 0.4999$; Figure A1). This diagnostic supports image defects as ordinal evidence for prioritization, but does not independently establish ordering among configurations of one generator. Configuration scores average per-observation predictions on a common cohort. Held-out permutation importance ranks L1 and LPIPS-t highest, coinciding with the strongest promoted associations (Appendix C).

Control-axis screening. Single-factor scoring retains three of seven controls: WorldDreamer guidance scale (0.609), temporal memory (0.538), and Cosmos front-view partial denoising (`FRONT_SIGMA_MAX`, 0.519). The remaining four score 0.320 or lower and are excluded from search; retention reflects local utility rather than range alone (Appendix C).

TPE candidate search and guardrail-aware ranking. Search completes five WorldDreamer rounds and 14 Cosmos rounds, and the resulting 19 round winners are frozen for full matched-frame confirmation.

5.5 Multi-Task Confirmation

The guardrail-aware score determines evaluation priority; the frozen evaluator stack determines each final decision. Nominated configurations are denoted WD-n and CS-n.

Table 2 distinguishes priority from acceptability. WD-1 reduces the composite gap by 0.1415 on the percentile-normalized scale (31.24% relative to its default) and has a positive tracking-gain interval. Its perception interval includes zero but meets the stated tolerance, and its planning estimate is near zero. CS-1 reduces the composite gap by 0.0952 (21.02% relative to its default), with positive nominal intervals on all three components. Relative reductions use each generator's default on the same

Table 2: Frame-level comparison relative to each default. Composite gains use the percentile-normalized scale; component gains use F1, HOTA, and meters. Intervals are nominal paired frame-bootstrap summaries from the held-out cohort and ignore within-scene dependence. Accept means passing the stated rule, not established scenario-level improvement; trade-off denotes an unresolved component condition. Bottom: remaining candidates.

Observed gain	WorldDreamer WD-1	Cosmos CS-1	Cosmos CS-2
Composite Γ_0 [95% CI]	$\mathbf{+0.1415}$ $[0.0908, 0.1922]$	$\mathbf{+0.0952}$ $[0.0556, 0.1348]$	-0.0452 $[-0.0953, 0.0049]$
Perception Γ_{per} [95% CI]	$+0.0052$ $[-0.0016, 0.0121]$	$+0.0481$ $[0.0386, 0.0576]$	$+0.0159$ $[0.0065, 0.0253]$
Tracking Γ_{trk} [95% CI]	$+0.0205$ $[0.0170, 0.0240]$	$+0.0098$ $[0.0053, 0.0142]$	-0.0029 $[-0.0090, 0.0031]$
Planning Γ_{plan} [95% CI]	$+0.0000$ $[-9 \times 10^{-6}, 9 \times 10^{-6}]$	$+0.0045$ $[0.0023, 0.0067]$	$+0.0053$ $[0.0026, 0.0080]$
Decision	**Accept**	**Accept**	Trade-off

Generator	Remaining candidates	Acceptance evidence	Rejects	Trade-offs
WorldDreamer	4	Aggregate reductions; component rule not met	0	4
Cosmos	12	No candidate satisfies the full rule	9	3
All confirmed	19	2 Accepts; 5 WD and 14 Cosmos rounds	9	8

confirmation observations, as defined in Appendix B. CS-2 improves the perception and planning gap estimates, but its tracking interval includes both improvement and deterioration and misses the zero-tolerance criterion. Its aggregate interval does not support improvement, so it is labeled a trade-off under the stated rule. Nine rejects and eight trade-offs show that ranking determines evaluation priority rather than acceptance.



Figure 3: Target-level recovery for WD-1. The default misses ID 4 at $T = 2$, whereas WD-1 retains it across all three displayed timesteps. Matched frames increase from 18 to 20, fragments decrease from 3 to 1, and the longest run increases from 14 to 20 frames. Solid and dashed boxes denote tracker outputs and aligned reference.

Qualitative confirmation case. Figure 3 illustrates WD-1's tracking output on one target: the default loses the identity at one timestep, whereas WD-1 keeps it. The case does not determine the acceptance decision; further cases are given in Appendix D.

Sampled-frame baseline comparison. Figure 4 compares one guided and one random search on the same fixed subset of eight dispersed frames. The composite gap on this subset is distinct from the confirmation-cohort gains in Table 2. The recorded runs reach final gaps of 0.447 and 0.453, with cumulative GPU allocations of 56.160 and 135.560 GPU·h for guided and random search, respectively. The comparison scope is discussed in Appendix E. With one run per method, the 0.006 difference is descriptive and does not establish a general end-to-end efficiency advantage.

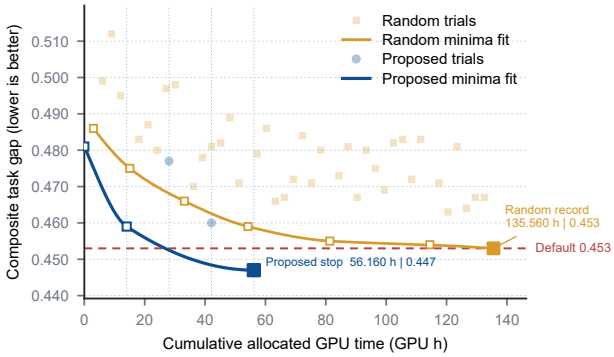

Figure 4: Search efficiency on the fixed eight-frame baseline (one run per method). Points show evaluated candidates, and curves are fits to the successive record-low gaps; the dashed line is the default configuration's gap on these frames.

6 Discussion and Limitations

Screening identifies WorldDreamer guidance scale and temporal memory, together with Cosmos front-view partial denoising, as representative high-impact native controls; this does not imply that the remaining parameters are ineffective, and control interactions are left to future work. The observed gains are agreement gains with respect to the fixed evaluator stack and may partly reflect adaptation to evaluator-specific characteristics; for the intended use, calibrating generated observations for evaluating that stack, this is the quantity of interest, but it is not evidence of better closed-loop driving.

Limitations. The evidence covers two pipelines, 26 InterHub scenes, three retained controls, and local control ranges, and no ablation isolates the contribution of the surrogate, the guardrails, or the control screen. The 12/6/8 scene-disjoint split avoids selection–confirmation leakage, but paired frame bootstrap ignores within-scene correlation: frames from one scenario are not independent, so the reported intervals are nominal frame-resampling intervals and the acceptance labels are decisions under the stated rule rather than evidence of scenario-level generalization. Scenario-level inference has not been performed, and only eight confirmation scenes are available. The search-efficiency comparison is a single run per method on eight frames. Future work will enlarge the confirmation set for scenario-based inference and repeat the search comparison across seeds.

7 Conclusion

We study inference-time configuration of frozen driving-video generators for the offline evaluation of autonomous-driving stacks. Image defects measured against condition-aligned real references provide evidence for prioritizing configurations, while downstream task-metric gaps determine the final decision. WD-1 and CS-1 meet the stated frame-level rule on held-out scenes, with composite-gap reductions of 31.24% and 21.02% relative to their respective defaults. These findings support an exploratory configuration workflow without generator retraining; they do not establish scenario-level statistical robustness or improved closed-loop driving performance.

AI USE STATEMENT

Generative AI assistance was used for manuscript revision, drafting and revising the evaluation-protocol description, preparation of the workflow illustration, and LaTeX formatting. The reported numerical results were retained from the existing experimental records; this assistance did not produce new generator experiments or replace the reported results with simulated outcomes.

ETHICS STATEMENT

This study concerns offline evaluation of generated driving observations using existing InterHub scenes and frozen models. Generated observations may omit traffic participants, distort motion, or reproduce biases in the source data and models. Improved agreement with the selected evaluator stack must therefore not be interpreted as evidence of safe real-world deployment or used as a substitute for safety validation. Reuse and sharing of driving data and generated outputs should respect dataset terms and protect identifiable road users. Further validation across diverse scenarios and evaluator stacks is needed before drawing conclusions about deployment safety.

REPRODUCIBILITY STATEMENT

Appendix A (Table A1) reports native control names and search ranges. Appendix B specifies defect preprocessing, surrogate aggregation, and evaluator settings, including the 12/6/8 scene-disjoint screening/search/confirmation protocol in Appendix B.4. Additional surrogate diagnostics, control-screening results, and qualitative cases are provided in Appendices C and D.

REFERENCES

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2546–2554, 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html.

A. J. Booker, J. E. Dennis, P. D. Frank, D. B. Serafini, V. Torczon, and M. W. Trosset. A rigorous framework for optimization of expensive functions by surrogates. *Structural Optimization*, 17(1): 1–13, 1999. doi: 10.1007/BF01197708.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020.

Francesca Campolongo, Jessica Cariboni, and Andrea Saltelli. An effective screening design for sensitivity analysis of large models. *Environmental Modelling & Software*, 22(10):1509–1518, 2007. doi: 10.1016/j.envsoft.2006.10.004.

Yun Chen, Frieda Rong, Shivam Duggal, Shenlong Wang, Xinchen Yan, Sivabalan Manivasagam, Shangjie Xue, Ersin Yumer, and Raquel Urtasun. GeoSim: Realistic video simulation via geometry-aware composition for self-driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7226–7236, 2021. doi: 10.1109/CVPR46437.2021.00715.

K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002. doi: 10.1109/4235.996017.

Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. MagicDrive: Street view generation with diverse 3D geometry control. In *International Conference on Learning Representations*, 2024a.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. VISTA: A generalizable driving world model with high fidelity and versatile controllability. In *Advances in Neural Information Processing Systems*, volume 37, pp. 91560–91596, 2024b. doi: 10.52202/079017-2906.

Jinwei Gu. Cosmos world foundation models for physical AI. In *Proceedings of the 3rd International Workshop on Rich Media With Generative AI*, pp. 39–39, 2025. doi: 10.1145/3746262.3761969.

Yanchen Guan, Haicheng Liao, Zhenning Li, Jia Hu, Runze Yuan, Guohui Zhang, and Chengzhong Xu. World models for autonomous driving: An initial survey. *IEEE Transactions on Intelligent Vehicles*, pp. 1–17, 2025. doi: 10.1109/TIV.2024.3398357.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Giles Hooker, Lucas Mentch, and Siyu Zhou. Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31(6):82, 2021. doi: 10.1007/s11222-021-10057-z.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2Drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In *Advances in Neural Information Processing Systems*, volume 37, pp. 819–844, 2024. doi: 10.52202/079017-0025.

Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. VAD: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8350, 2023.

Xiyan Jiang, Xiaocong Zhao, Yiru Liu, Zirui Li, Peng Hang, Lu Xiong, and Jian Sun. A naturalistic trajectory dataset with dense interaction for autonomous driving. *Scientific Data*, 12(1):1084, 2025. doi: 10.1038/s41597-025-05344-7.

Napat Karnchanachari, Dimitris Geromichalos, Kok Seang Tan, Nanxiang Li, Christopher Eriksen, Shakiba Yaghoubi, Noushin Mehdipour, Gianmarco Bernasconi, Whye Kit Fong, Yiluan Guo, and Holger Caesar. Towards learning-based planning: The nuPlan benchmark for real-world autonomous driving. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 629–636, 2024. doi: 10.1109/ICRA57147.2024.10610077.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European Conference on Computer Vision*, volume 13669 of *Lecture Notes in Computer Science*, pp. 1–18, 2022. doi: 10.1007/978-3-031-20077-9_1.

Carl Lindström, Georg Hess, Adam Lilja, Maryam Fatemi, Lars Hammarstrand, Christoffer Petersson, and Lennart Svensson. Are NeRFs ready for autonomous driving? towards closing the real-to-simulation gap. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4461–4471, 2024. doi: 10.1109/CVPRW63382.2024.00449.

William Ljungbergh, Adam Tonderski, Joakim Johnander, Holger Caesar, Kalle Åström, Michael Felsberg, and Christoffer Petersson. NeuroNCAP: Photorealistic closed-loop safety testing for autonomous driving. In *Computer Vision – ECCV 2024*, pp. 161–177, 2024. doi: 10.1007/978-3-031-73404-5_10.

Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang, and Li Zhang. WoVoGen: World volume-aware diffusion for controllable multi-camera driving scene generation. In *Computer Vision – ECCV 2024*, pp. 329–345, 2024. doi: 10.1007/978-3-031-72989-8_19.

Samarth Mishra, Rameswar Panda, Cheng Perng Phoo, Chun-Fu Richard Chen, Leonid Karlinsky, Kate Saenko, Venkatesh Saligrama, and Rogerio S. Feris. Task2Sim: Towards effective pre-training and transfer from synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9194–9204, 2022. doi: 10.1109/CVPR52688.2022.00898.

Max D. Morris. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174, 1991. doi: 10.1080/00401706.1991.10484804.

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

Galit Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010. doi: 10.1214/10-STS330.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):307, 2008. doi: 10.1186/1471-2105-9-307.

Adam Tonderski, Carl Lindström, Georg Hess, William Ljungbergh, Lennart Svensson, and Christoffer Petersson. NeuRAD: Neural rendering for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14895–14904, 2024. doi: 10.1109/CVPR52733.2024.01411.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Jiahao Wang, Zhenpei Yang, Yijing Bai, Yingwei Li, Yuliang Zou, Bo Sun, Abhijit Kundu, Jose Lezama, Luna Yue Huang, Zehao Zhu, Jyh-Jing Hwang, Dragomir Anguelov, Mingxing Tan, and Chiyu Max Jiang. Drive&Gen: Co-evaluating end-to-end driving and video generation models. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 8934–8941, 2025. doi: 10.1109/IROS60139.2025.11247678.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. DriveDreamer: Towards real-world-drive world models for autonomous driving. In *Computer Vision – ECCV 2024*, pp. 55–72, 2024a. doi: 10.1007/978-3-031-73195-2_4.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14749–14759, 2024b. doi: 10.1109/CVPR52733.2024.01397.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6902–6912, 2024. doi: 10.1109/CVPR52733.2024.00659.

Tianyi Yan, Dongming Wu, Wencheng Han, Junpeng Jiang, Xia Zhou, Kun Zhan, Cheng-Zhong Xu, and Jianbing Shen. DrivingSphere: Building a high-fidelity 4D world for closed-loop simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27531–27541, 2025. doi: 10.1109/CVPR52734.2025.02564.

Xuemeng Yang, Licheng Wen, Tiantian Wei, Yukai Ma, Jianbiao Mei, Xin Li, Wenjie Lei, Daocheng Fu, Pinlong Cai, Min Dou, Liang He, Yong Liu, Botian Shi, and Yu Qiao. DriveArena: A closed-loop generative simulation platform for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26933–26943, 2025. doi: 10.1109/ICCV51701.2025.02500.

Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. UniSim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1389–1399, 2023. doi: 10.1109/CVPR52729.2023.00140.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, volume 13682 of *Lecture Notes in Computer Science*, pp. 1–21, 2022. doi: 10.1007/978-3-031-20047-2_1.

Guosheng Zhao, Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Xueyang Zhang, Yida Wang, Guan Huang, Xinze Chen, Boyuan Wang, Youyi Zhang, Wenjun Mei, and Xingang Wang. DriveDreamer4D: World models are effective data machines for 4D driving scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12015–12026, 2025. doi: 10.1109/CVPR52734.2025.01122.

APPENDICES

A NATIVE INFERENCE-TIME CONTROLS AND SEARCH RANGES

Table A1 lists the seven native controls included in single-factor screening and their specified ranges. These controls provide complementary coverage of guidance and conditioning strength, temporal memory, foreground weighting, and front-view denoising across the two pipelines. Each can be varied individually with fixed model weights, making this a focused, tractable screen of representative inference-time mechanisms rather than an exhaustive study of every exposed parameter. Screening determines which of these seven axes are retained for local search; other native controls remain outside this study.

Table A1: The seven native inference-time controls included in single-factor screening and their specified search ranges. Screening determines the retained axes for local search.

Generator	Control parameter	Default	Search range
WorldDreamer	`controlnet_conditioning_scale`	1.0	[0.5, 1.5]
	`temporal_memory_strength`	1.0	[0.3, 1.0]
	`guidance_scale`	2.0	[1.01, 4.0]
	`fg_weight`	1.0	[0.5, 1.5]
Cosmos	`FRONT_SIGMA_MAX`	95	[60, 95]
	`MV_GUIDANCE`	3.0	[1.5, 4.5]
	`hdmap_control_weight`	1.0	[0.47, 1.0]

B METHOD IMPLEMENTATION DETAILS

B.1 DEFECT AND TASK-GAP PREPROCESSING

The paired audit contains 864 aligned image-defect observations, 432 per generator, sampled at fixed temporal intervals in 26 interaction-rich InterHub scenes. In this study, real image references are retrieved from the nuScenes scenes corresponding to the selected InterHub trajectories (Caesar et al., 2020). Each observation is aligned by scene, timestamp, and camera view. Scenes define the unit for data partitioning, while individual observations provide the evaluation samples. For every condition-aligned observation, the six-dimensional defect vector contains L1 error, direction-aligned PSNR, LPIPS, projection error, LPIPS-t, and optical-flow error. PSNR is multiplied by -1 before modeling so that every channel increases with poorer quality. Temporal features at $t = 0$ and unavailable projection values are imputed using the corresponding training fold only.

Abbreviations and notation. PSNR denotes peak signal-to-noise ratio; RBF, radial basis function; SVR, support vector regression; MAE, mean absolute error; RMSE, root mean squared error; and IoU, intersection over union. LPIPS denotes Learned Perceptual Image Patch Similarity, LPIPS-t its temporal variant, and BH the Benjamini–Hochberg correction. CI denotes confidence interval, TPE tree-structured Parzen estimator, and HD map high-definition map. Conditions are indexed as c_i, control axes by p, repeated random seeds by $s \in \mathcal{R}$, and search rounds by r. The shorthand $p = v$ means setting control coordinate θ_p to v, with all other coordinates fixed. Thus c_i is not a control index, and s is distinct from the round index in $\theta_m^{(r)}$.

Perception, tracking, and planning fidelity gaps are the absolute real–generated differences in F1, HOTA, and planning trajectory L_2, respectively. They are normalized independently using the training-fold 5th and 95th percentiles, clipped to $[0, 1]$, and averaged with equal weights $\omega_k = 1/3$ to form the composite endpoint. No candidate-set statistic is used for normalization. For the RQ1 association analysis, ranked defect and gap variables are residualized with additive generator and scene effects. Benjamini–Hochberg correction is applied jointly across tested paths, and a promoted path must retain a positive sign within both generator pipelines.

Explicitly, for frozen reference percentiles $p_{5,k}$ and $p_{95,k}$,

$$\widetilde{G}_{i,k} = \mathrm{clip}_{[0,1]}\left(\frac{G_{i,k} - p_{5,k}}{p_{95,k} - p_{5,k}}\right).$$

Throughout the paper, the percentile-normalized composite scale refers to this clipped, training-reference percentile scale; the frozen reference values used outside cross-validation are reported in Table A2. Cross-validation recomputes these quantities from each training fold. Composite gains are differences of means on that scale, not standard-deviation units or relative percentages of the default gap. Task-component gains retain F1, HOTA, and meter units. Default and candidate comparisons must use the same valid observations and frozen percentiles; audit means and confirmation gains from different valid subsets cannot simply be subtracted.

Table A2: Frozen reference percentiles for absolute task gaps. Planning values are in meters; other values are dimensionless.

Gap	p_5	p_{95}
Perception	0.055803177815558665	0.5282330827067667
Tracking	0.10085030732339233	0.3043478260869565
Planning	0.16284932933617993	7.842150497113326

Relative composite-gap reductions are reported separately as $100\,\Gamma_0(\theta)/A(\theta_0)$, using each generator's default on the same valid confirmation observations and frozen normalization.

B.2 Control Screening, Surrogate, and Candidate Ranking

Each control axis is first swept independently with all other controls fixed at the default and with matched scenes and seeds. The RBF-SVR takes the absolute, direction-aligned defect vector for each observation, using the same fitted imputation and feature standardization in training and inference. Let f include this frozen preprocessing. The predicted default-relative change is

$$\Delta\widehat{a}_p(v) = \frac{1}{|\mathcal{I}||\mathcal{R}|}\sum_{i,s}\left[f(\mathbf{z}_{i,s}(p{=}v)) - f(\mathbf{z}_{i,s}(\theta_0))\right].$$

The default-relative defect vector describes response mobility; it is not substituted for the absolute SVR input. Predictions are made before averaging because the RBF mapping is nonlinear. The axis score combines the magnitude of the predicted change with scene–seed response stability. For scene–seed group b, let d_b be the mean default-relative predicted change and let d be its observation-weighted aggregate. Stability is the fraction of groups for which $\mathrm{sign}(d_b) = \mathrm{sign}(d)$; it is zero when d is numerically close to zero. The magnitude transform is $\mathcal{N}_{\mathrm{ref}}(|d|) = \min(|d|/s_{\mathrm{ref}}, 1)$. Each scanned value receives the product of this magnitude and stability; the control-axis score is the maximum across its scanned values. Promoted-channel mobility is scaled by the frozen defect-reference $p_{95} - p_5$ range. Retention requires both an axis score at least $S_{\min}$ and a maximum absolute promoted-channel movement at least $M_{\min}$. These maxima are taken over the scanned values and need not occur at the same value. The reference implementation defaults are $S_{\min} = 0.1$, $M_{\min} = 0.02$, and $s_{\mathrm{ref}} = 1.0$. Retained axes define bounded continuous domains for TPE search.

The defect-association surrogate is an RBF-kernel support vector regressor whose output is the predicted normalized composite task gap. Missing-value imputation, feature standardization, and task-gap normalization are fitted within each training fold. Five-fold GroupKFold keeps all frames from one scene in the same fold, and kernel hyperparameters are selected by an inner scene-grouped split. The implementation uses median imputation followed by feature standardization, with five outer and three inner scene-grouped folds. Its candidate grid is $C \in \{0.3, 1, 3, 10\}$, $\epsilon \in \{0.01, 0.05\}$, and $\gamma \in \{\texttt{scale}, 0.1, 0.3\}$, with an RBF kernel. Selection maximizes inner-fold mean Spearman correlation, breaking ties by lower mean MAE. Here `scale` denotes the inverse of the feature count times the training-input variance. This grid specifies the selection procedure rather than a single final SVR configuration. After single-factor screening, a local tree-structured Parzen estimator operates on the retained continuous domain (Bergstra et al., 2011). At each round, TPE generates five configurations around the current sampling center. For retained axis p with global bounds $[\ell_p, u_p]$, center h_p, and radius b_p, the local interval is $[\max(\ell_p, h_p - b_p), \min(u_p, h_p + b_p)]$; the multi-axis domain

is their Cartesian product. When no radius is supplied, the implementation sets $b_p = 0.25(u_p - \ell_p)$. All five candidates are scored by R, and each score is added to the TPE proxy history. Only the lowest-R candidate receives downstream evaluation. The other four remain in the proxy history without downstream outcomes for that round. Predictions are averaged over the same search-cohort observations for each configuration, then combined with one-sided penalties when promoted RQ1 channels deteriorate beyond their guardrails. The guardrail uses $\lambda = 0.25$, $\tau_j = 0$, and training-cohort 5th–95th percentile ranges for normalizing excess. The lowest-scoring configuration is evaluated on all frames in the experimental search cohort. If its observed downstream composite gap improves on the current configuration, it becomes both the retained best configuration and the next-round sampling center; otherwise both remain unchanged. The eight dispersed frames are used only for the separate random-search efficiency baseline. Search settings and candidate ordering are frozen before final confirmation results are inspected.

B.3 Native Controls and Conditioning Scope

WorldDreamer uses trajectory and road-condition inputs together with a fixed built-in style reference. Its recorded checkpoint is `dreamer_pretrained/SDv1.5_mv_single_ref_nus/weight_S200000` within the DriveArena WorldDreamer directory. The default configuration specifies `ref_images=builtin_boston_style` and `use_nuplan_interhub_calib=false`. Cosmos front-view generation receives text and an HD-map control video through the AV-sample pathway, using `nvidia/Cosmos-Transfer1-7B-Sample-AV/base_model.pt` and `nvidia/Cosmos-Transfer1-7B-Sample-AV/hdmap_control.pt`. The remaining views use the `nvidia/Cosmos-Transfer1-7B-Sample-AV-Single2MultiView` checkpoint family, with the generated front-view anchor supplied through `view_condition_video` alongside multi-view control conditions. Paired real evaluation frames are not supplied as scenario-specific reference images. Trajectory and road-condition inputs, seeds, and evaluation alignment are held fixed between the default and candidate configurations. L1 and PSNR compare generated observations with their aligned real evaluation references, not the fixed style reference.

Black-box configuration means varying exposed native controls with fixed architecture and weights; it does not involve adding trainable parameters. Appendix A reports native parameter names and ranges. Table A3 lists the seven controls examined by single-factor screening and their outcomes. The retained search domain has two dimensions for WorldDreamer and one for Cosmos.

Table A3: Native controls and single-factor screening scores. Front partial denoising denotes `FRONT_SIGMA_MAX`, and conditioning scale denotes `controlnet_conditioning_scale` (labeled "condition control" in Figure A2). Retention follows the score and promoted-channel criteria in Section 3.3.

Pipeline	Native control	Score	Retain
WorldDreamer	Guidance scale	0.609	Yes
WorldDreamer	Temporal memory	0.538	Yes
WorldDreamer	Conditioning scale	0.320	No
WorldDreamer	Foreground weight	0.098	No
Cosmos	Front partial denoising	0.519	Yes
Cosmos	HD-map control	0.061	No
Cosmos	Multi-view guidance	0.025	No

B.4 Scene-Disjoint Evaluation

The 26 scenes are partitioned into 12 screening, six search, and eight confirmation scenes. Each scene is assigned as a unit: all frames, cameras, seeds, and both generator versions share its assignment. The three cohorts are scene-disjoint, and cohort sizes count scenes rather than frames. The reported statistics are obtained under this 12/6/8 protocol.

Within the 12 screening groups, five-fold scene-grouped validation fits all preprocessing and surrogate hyperparameters without access to search or confirmation outcomes. The fitted model, normalization, guardrails, and margins are then frozen. Search uses the six search groups and budgets of

five WorldDreamer rounds and 14 Cosmos rounds, with five proposals per round. Candidate identities and their order are frozen before evaluation on the eight confirmation groups; confirmation outcomes do not update the surrogate or the search center.

Confirmation compares each nominated candidate with its generator default on the eight held-out groups. For each task, default and candidate use matched valid observations. The reported confidence intervals use paired frame bootstrap as specified below, with task and composite gains computed using the frozen screening normalization. Scene-disjoint cohort assignment separates candidate selection from confirmation; the bootstrap resampling unit remains the paired frame.

B.5 Frozen Evaluators and Confirmation

Perception uses BEVFormer with the R101-DCN 24-epoch checkpoint `bevformer_r101_dcn_24ep.pth` (Li et al., 2022). It processes up to 48 frames from six cameras, uses a detection threshold of 0.2, and matches visible objects within 40 m at a 3-m distance threshold. Perception references are constructed from non-ego object positions in `agent_data_dt.feather`, transformed into ego coordinates using the corresponding ego pose before matching predictions. Tracking references use the scene trajectories, object identities, and dimensions to project ground-truth boxes into each camera view. Matched real, default, and candidate records share the same ground-truth scene directory and timestamp mapping. Real-video model predictions are not used as ground truth. Tracking uses YOLOv8n detections followed by ByteTrack association (Zhang et al., 2022), with detector confidence 0.25, detector IoU setting 0.5, ground-truth matching IoU 0.3, and `yolov8n.pt` weights. These settings are shared across both generators and across their real, default, and candidate records. Planning uses VAD configuration `VAD_tiny_stage_2.py` and checkpoint `VAD_tiny.pth` (Jiang et al., 2023). The planning reference is the ego trajectory from `agent_data_dt.feather`, transformed using the current ego position and heading. Predicted and reference displacement sequences are accumulated into positions; the error averages two-dimensional Euclidean distances over up to six valid future points, rather than measuring only the endpoint. Ego-state inputs include position, quaternion, velocity, heading, and trajectory history. The auxiliary vector is $[v_x, v_y, 0, 0, 0, L_{\text{ego}}, W_{\text{ego}}, \text{speed}, 0]$, where L_{ego} and W_{ego} are vehicle length and width.

HOTA is computed over the temporal window associated with each sampled timestamp, using corresponding windows for real, default, and candidate sequences. Thus $q_{\text{trk}}(x_i)$ denotes a window-level tracking score indexed by observation i, not a metric evaluated from a single image. Each tracking record used in paired comparisons represents its associated temporal window.

The same preprocessing and evaluator stack are applied to aligned real observations, generator defaults, and nominated configurations. Candidate task-fidelity gains are summarized by paired frame bootstrap, with positive values denoting smaller absolute gaps than the corresponding generator default. Each resample carries the aligned real, default, and candidate outputs together; unavailable planning outputs are excluded from the corresponding planning and composite comparison rather than imputed. Task-specific non-inferiority margins are the maximum scene-wise 95th-percentile absolute frame differences from repeated identical-input real–real evaluator runs. On three real scenes comprising 144 frames, the frozen evaluators span F1 $[0.000, 0.788]$, HOTA $[0.171, 0.303]$, and planner L_2 $[9.56, 32.93]$ m; the resulting margins are 0.0452 F1, 0.0000 HOTA, and 0.0031 m. These margins describe evaluator repeatability and serve as component-wise decision tolerances, rather than real-world driving-performance thresholds. Under the stated rule, the nominal lower bound on a component gain must be at least $-\mu_k$; zero HOTA tolerance therefore requires a nonnegative lower bound. These frame-resampling summaries do not account for within-scene or within-window dependence.

C DEFECT-ASSOCIATION, SURROGATE, AND CONTROL-SCREENING DIAGNOSTICS

Table A4: Promoted residual-rank associations. All paths survive false-discovery correction and remain positive within both generators.

Defect	Task	ρ	BH q
L1 reconstruction	Planning	0.548	4.1×10^{-9}
Aligned PSNR	Planning	0.522	2.0×10^{-8}
LPIPS-t	Planning	0.502	6.5×10^{-8}
L1 reconstruction	Perception	0.366	2.6×10^{-4}
Aligned PSNR	Perception	0.341	7.0×10^{-4}
LPIPS-t	Perception	0.320	1.5×10^{-3}
Flow error	Planning	0.281	5.9×10^{-3}

The scene-held-out surrogate diagnostic is shown in Figure A1. The pooled diagnostic has Spearman $\rho = 0.8338$ and moderate exact-value calibration ($R^2 = 0.4999$). It provides evidence of ordinal information in image-defect measurements, but does not independently validate configuration-level ranking. Held-out permutation importance ranks L1 and LPIPS-t highest, consistent with the promoted reconstruction and temporal associations. Projection error and LPIPS rank above flow error and direction-aligned PSNR without being promoted associations.

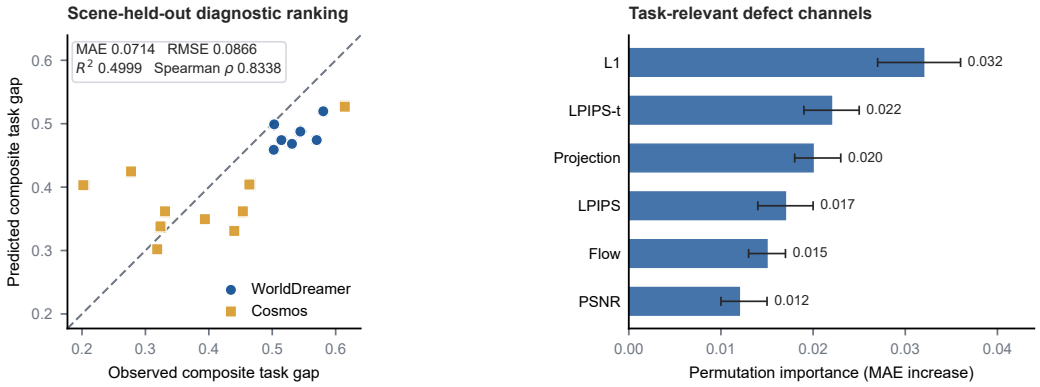

Figure A1: Scene-held-out surrogate performance. Predicted and measured composite gaps have MAE 0.0714, RMSE 0.0866, $R^2 = 0.4999$, and Spearman $\rho = 0.8338$. Held-out permutation importance ranks L1 and LPIPS-t highest.

Figure A3 summarizes how the retained controls connect to the promoted defect channels and task gaps, and Figure A2 reports the complete single-factor evidence. The score profile combines predicted task-gap movement with response stability, while the mobility summary reports default-relative defect responses and their association with the surrogate ranking score. WorldDreamer guidance scale, WorldDreamer temporal memory, and Cosmos front-view partial denoising show strong movement in task-associated channels and are retained; the remaining four axes have weaker or less stable responses within the tested ranges and are not included in the current local combination search. This screening decision does not establish their ineffectiveness under other ranges or interactions.

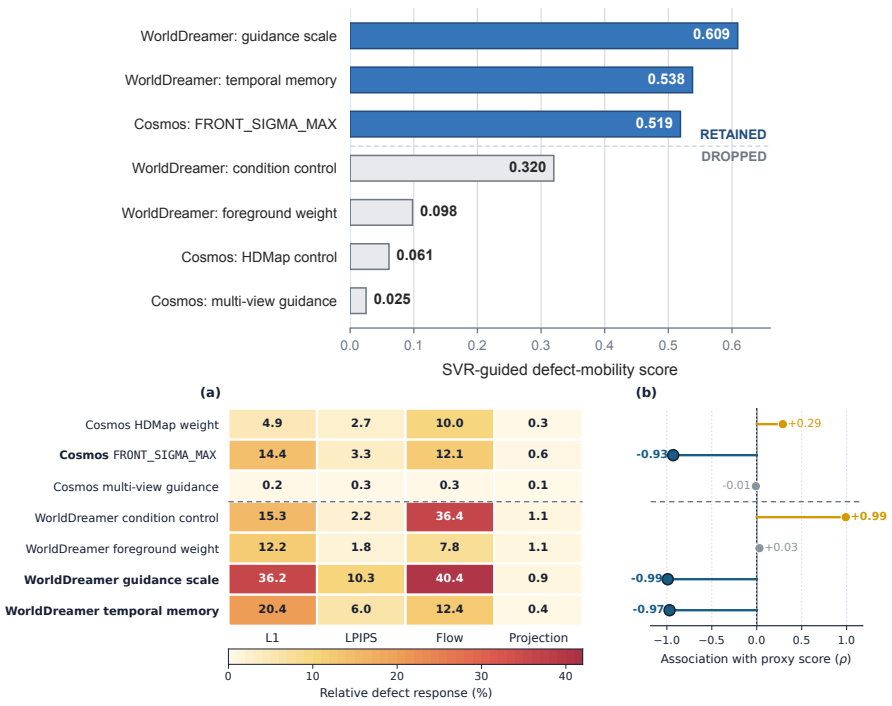

Figure A2: Complete single-factor screening evidence. Top: predicted task-gap movement and response stability retain three controls for local search. Bottom left: default-relative response ranges for L1, LPIPS, flow, and projection defects. Bottom right: Spearman association between each control response and the surrogate ranking score. Bold labels and outlined markers identify the retained axes.

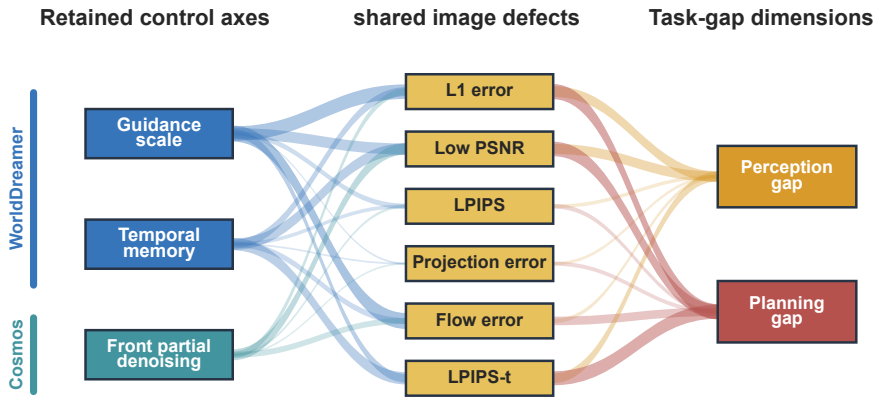

Figure A3: Control-axis screening evidence. Left bands encode control-induced defect mobility; right bands distinguish the seven promoted associations in Table A4 from thin qualitative links ("Low PSNR" denotes direction-aligned PSNR). The tracking gap has no promoted association and is therefore absent on the right. All six defects remain surrogate inputs; screening reduces only the control search space.

D Additional Decision Cases

The main text reports the frame-level confirmation statistics and one target-level tracking-recovery case. This section adds the qualitative six-view illustration of the accepted configuration and one trade-off case.

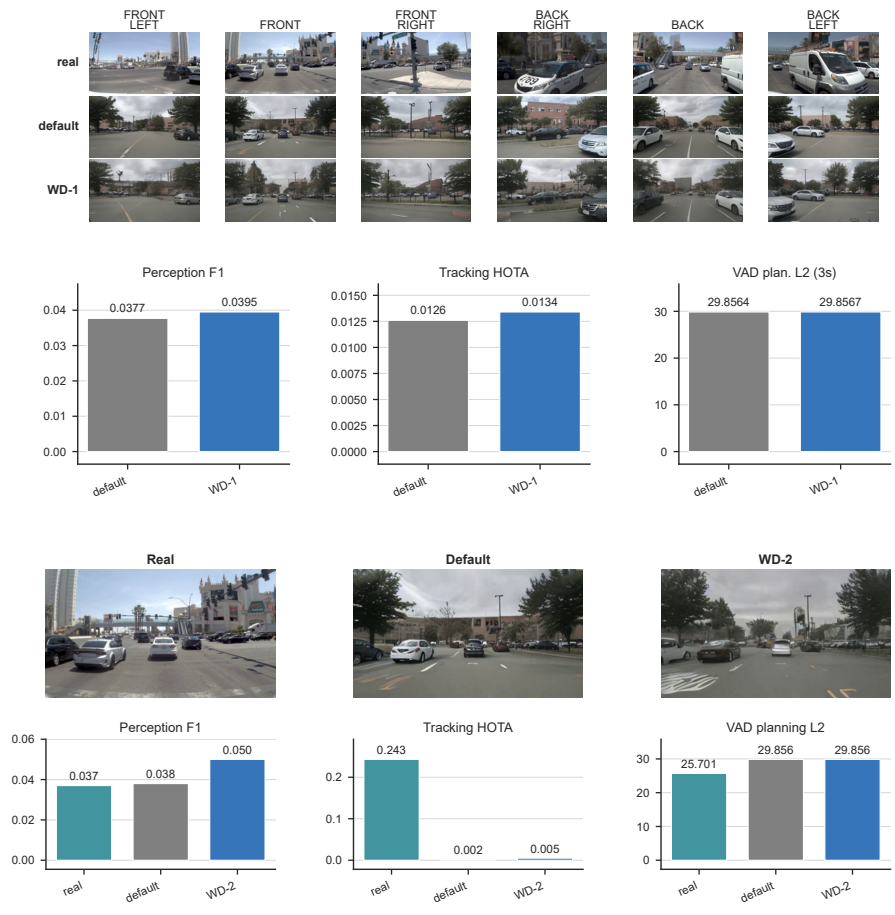

Figure A4: Additional WorldDreamer decision cases. Top: task-level illustration of configuration WD-1 under matched conditions, showing raw task scores for the default and candidate. Bottom: candidate WD-2 on scene `4bb1341b4ab85bc6` raises the raw perception F1 above the real reference, which *widens* the perception gap while narrowing the tracking gap and leaving the planner error unchanged. The examples illustrate task outputs; cohort-level decisions use the stated rule.

D.1 The Accepted Configuration on a Matched Six-View Case

The top panel of Figure A4 illustrates configuration WD-1 on a matched six-view case. Relative to the default under the same conditions, raw perception F1 rises from 0.0377 to 0.0395, and raw tracking HOTA rises from 0.0126 to 0.0134. The planning L_2 change of 0.0003 m remains below the 0.0031-m repeatability margin. This example illustrates task outputs and does not determine the cohort-level decision.

D.2 The Accepted Cosmos Configuration

The accepted Cosmos configuration CS-1 lies on the retained front-view partial-denoising axis. Relative to the Cosmos default, it yields a composite task-fidelity gain of 0.0952 on the percentile-normalized scale (nominal 95% frame-bootstrap interval, 0.0556–0.1348), with positive gains of 0.0481 F1, 0.0098 HOTA, and 0.0045 m for perception, tracking, and planning, respectively. Its composite gain and component-wise tolerances satisfy the same frame-level acceptance rule used for WD-1.

D.3 A Task Trade-Off

The bottom panel of Figure A4 illustrates the kind of conflict that the trade-off label records at one sampled timestamp, with tracking evaluated over its associated temporal window. On scene `4bb1341b4ab85bc6`, candidate WD-2 raises the raw perception F1 from 0.038 to 0.050 while the real reference scores 0.037; the perception gap therefore widens from 0.001 to 0.013. At the same time, the tracking gap narrows (HOTA 0.002 to 0.005 against 0.243) and the planner error is unchanged. A higher raw score is thus not higher fidelity: the objective is agreement with the real reference, not a larger metric. At the cohort level, WD-2 is labeled a trade-off because the nominal lower bound on its perception-gap gain falls below the stated tolerance, while another component has a positive nominal gain lower bound. This case illustrates why confirmation uses task-component checks in addition to the overall aggregate.

E Scope of the Search-Efficiency Comparison

The eight-frame comparison reports one search trajectory per method. The 45 random-search candidates are not 45 independent search runs. Recorded cumulative GPU allocations are 56.160 and 135.560 GPU·h for guided and random search, respectively. No across-run uncertainty or component-level cost breakdown is reported. These single-run results describe the recorded comparison and do not establish a general end-to-end efficiency advantage.

F Interpretation Scope of Transferred Evaluators

The absolute levels of the downstream metrics should be interpreted in the context of evaluator transfer to InterHub rather than as direct estimates of deployment performance. Perception, tracking, and planning evaluators may be imperfectly calibrated outside their original evaluation distributions. Accordingly, no absolute outputs are compared across different evaluator architectures or datasets. Each generated observation and its aligned real reference are instead processed by the same frozen evaluator under matched scene, map, ego trajectory, camera view, and temporal conditions. Repeated real–real evaluations estimate evaluator-specific repeatability, and their 95th-percentile absolute differences define the component-wise non-inferiority margins used during confirmation. These margins are not subtracted from the reported task gaps. The resulting analysis measures paired changes associated with replacing an aligned real observation with its generated counterpart under a fixed evaluator stack, rather than absolute driving capability or real-world deployment performance.