TWINWORLD: CONTRAST-CONSISTENT CAUSAL RESPONSE-SURFACE LEARNING FOR REALITY–BELIEF COUNTERFACTUALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable counterfactual modeling requires distinguishing changes in the physical world from changes in an agent's belief about that world. Under partial observability, the same behavioral response can arise from different underlying causes, making conventional intervention-based evaluation insufficient to identify whether behavior changes are driven by physical conditions, beliefs, or their interaction. We formulate this challenge as a causal response-surface learning problem and introduce a paired factorial intervention framework that independently manipulates physical state and initial belief across four counterfactual worlds, $(P, K) \in \{(0,0), (1,0), (0,1), (1,1)\}$. This design identifies three interpretable causal contrasts: the physical main effect, the belief main effect, and their interaction.

Building on this formulation, we develop TWINWORLD, a contrast-consistent response-surface estimator that jointly recovers physical openness, belief openness, and target relevance, and reconstructs the four potential-response cells through a shared structured decoder with a bounded residual. The resulting factorization couples latent-state prediction with intervention algebra, providing an explicit inductive bias for counterfactual consistency rather than treating intervention conditions as unrelated outcomes. We further characterize the identifiability of belief from behavioral evidence, showing that belief-independent actions provide no information about the hidden belief state, whereas repeated observations can improve recoverability under persistent beliefs and conditional independence.

We evaluate the framework using controlled interventions and semi-synthetic traffic environments with map and motion-history evidence. The experiments examine route-response estimation, latent-state recovery, intervention contrasts, and negative controls, demonstrating that temporal evidence supports hidden-state recovery and that contrast-consistent structure improves counterfactual response estimation. Overall, TWINWORLD provides an operational causal framework for separating reality from belief and for learning structured counterfactual response surfaces when latent beliefs mediate observed behavior.

1 INTRODUCTION

A traffic simulator may reproduce a factual detour while answering the wrong counterfactual question. A driver can leave a route because a road is physically closed, because an incorrect message makes the driver believe that it is closed, or because both conditions hold. These mechanisms can produce the same binary route outcome while implying different intervention semantics. A representation that collapses physical state and belief state can therefore achieve good factual prediction without exposing which intervention produced the response.

We study this problem through a controlled structural decomposition. Let $P \in \{0, 1\}$ denote whether a target link is physically closed and let $K \in \{0, 1\}$ denote whether the agent initially believes that the link is closed. We use the convention $P = 0$ for physically open, $P = 1$ for physically closed, $K = 0$ for belief-open, and $K = 1$ for belief-closed. For a fixed source scene, target link, 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.

policy, and random seed, the four potential responses form the response surface

$$\mathbf{Y}_s = \big(Y_s(0,0), Y_s(1,0), Y_s(0,1), Y_s(1,1)\big). \tag{1}$$

This factorial construction separates the physical main effect, belief main effect, and non-additive interaction without requiring a continuous counterfactual trajectory to be observed.

The central methodological point is to make the causal estimand executable and the response surface structurally constrained. Our benchmark defines $do(P = p, K = k)$ directly on the graph-and-belief state and obtains the reference response from a deterministic shortest-path oracle. This separates intervention semantics from outcome modeling and avoids circular evaluation against a learned planner. The resulting object is a route-level factorial response surface for a specified policy, which lets us evaluate causal contrasts directly rather than infer them indirectly from factual trajectory error.

This formulation also separates three questions that are often conflated. First, can the representation recover the manipulated physical and belief states? Second, can it predict the binary route response? Third, can it recover the signed factorial contrasts between intervention cells? We report all three rather than reducing them to a single trajectory metric.

Our contributions are: (i) a four-cell reality–belief causal design in which physical, epistemic, and interaction effects are explicit factorial estimands; (ii) a *contrast-consistent response-surface estimator* that couples all four intervention cells through target relevance and two interpretable openness coordinates, reducing an unconstrained four-arm outcome model to a structured three-coordinate representation; (iii) a paired causal objective and bounded decoder correction with an analytic error-propagation result from latent-state estimation to factorial contrast error; (iv) an executable graph-intervention benchmark with an independently defined route oracle, enabling direct evaluation of intervention effects; and (v) a multi-axis empirical study combining state recovery, route-response accuracy, negative controls, capacity-matched causal baselines, seed variation, and controlled cost shift. Together these components turn the four-cell intervention grid into a learnable causal response surface whose representation, estimands, and error budget are explicitly linked.

Why factorization is a causal-estimation contribution. An unconstrained four-arm predictor treats $(Y_{00}, Y_{10}, Y_{01}, Y_{11})$ as unrelated treatment-specific outcomes. TWINWORLD instead starts from the intervention-consistent surface $r[0, p, k, p + k - pk]$ and learns only three interpretable coordinates (r, π_P, π_K) before a bounded correction. This is a structural regularizer on the counterfactual response surface: it removes degrees of freedom that cannot be attributed to the specified intervention semantics, while retaining a controlled residual channel for model misspecification. Consequently, the same latent variables define both the causal representation and the factorial contrasts, and Proposition 1 converts latent-state estimation error into an explicit counterfactual contrast-error budget.

Key empirical evidence. The experiments probe the proposed causal structure at several levels. First, TWINWORLD reaches route-response F1 0.7038, the strongest valid matched configuration, while simultaneously recovering physical state, belief state, and target relevance. Second, removing the relevance gate changes both route F1 and CEE sharply, providing a mechanism-level ablation rather than only an end-point comparison. Third, the controlled cost shift yields CEE 0.1148, outperforming direct structural decoding (0.1742) and TARNet-4 (0.3299) in that shifted environment. Together with the factorial contrast table and aggregate AIPW/DR calibration, these results test whether the proposed factorization behaves as a causal response-surface estimator rather than merely as a route classifier.

Reading Figure 1. The open circles labeled O and D mark the origin and destination of the fixed route query. The dashed path is the reference route, while the solid blue path is the executed route under the intervention. In the physical-closure panel, the solid vertical barrier denotes a link made physically unavailable ($P = 1$). In the belief-closure panel, the diamond marks the agent's perceived barrier, while the speech-bubble symbol denotes the belief-side information cue; the underlying link remains physically open ($P = 0$). The arrows between panels indicate that the two interventions share the same scene and route query but differ in the manipulated layer. The right-hand panel records the binary route response $Y = 1$, so the figure illustrates intervention semantics rather than an observed human trajectory.

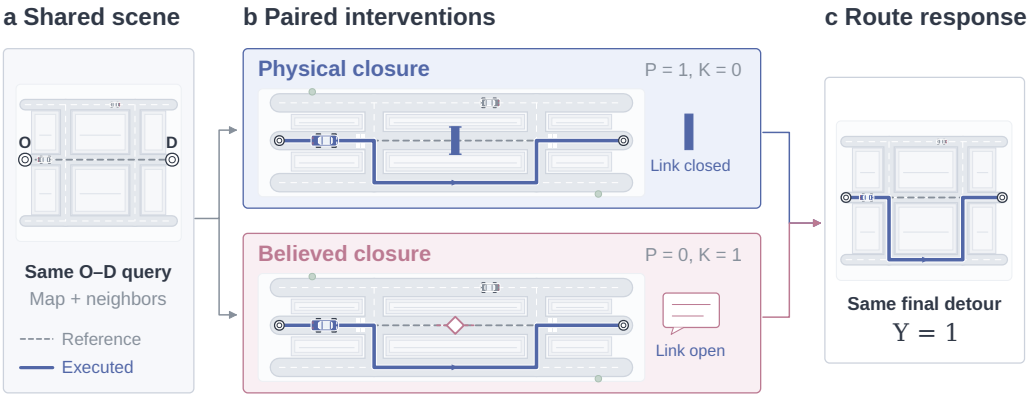

Figure 1: Reality–belief separation. The same route query admits distinct physical and belief interventions with an identical final route departure.

2 Related Work

Traffic simulation and forecasting. Microscopic simulators model vehicle dynamics and interactions, while data-driven simulators learn or replay behavior from logged scenes (Alvarez Lopez et al., 2018; Gulino et al., 2023). Argoverse 2 provides map-grounded motion scenarios and HD maps (Wilson et al., 2023); GameFormer studies interactive prediction and planning (Huang et al., 2023); SceneDiffuser++ targets generative city-scale traffic simulation (Tan et al., 2025). These systems are complementary to the present task: our estimand is a controlled route response under paired physical and epistemic interventions.

Causal representation learning. Potential-outcome and structural-causal frameworks formalize intervention-specific counterfactuals (Rubin, 1974; Pearl, 2009). TARNet/CFR, meta-learners, DragonNet, and DRNet provide treatment-effect parameterizations (Shalit et al., 2017; Künzel et al., 2019; Shi et al., 2019; Schwab et al., 2020), while doubly robust estimators target aggregate effects (Bang and Robins, 2005; Chernozhukov et al., 2018). These methods typically learn treatment-specific outcomes or aggregate contrasts; our distinction is that the four intervention cells are coupled by a known response-surface algebra before the outcome is decoded. The benchmark is fully crossed, so the comparison isolates response parameterization rather than treatment-assignment correction.

Interactive simulators versus counterfactual estimators. Waymax, GameFormer, and SceneDiffuser++ address simulation, interactive prediction, or generative rollout (Gulino et al., 2023; Huang et al., 2023; Tan et al., 2025). They are important application-level references but do not directly define the paired physical-versus-belief factorial estimand studied here. We therefore use them to delimit the problem rather than claiming an apples-to-apples leaderboard comparison.

Invariance and transportability. Invariant prediction asks whether causal relations persist across environments (Peters et al., 2016). Our cost-shift experiment is a within-family stress test of that idea. It is deliberately interpreted as environment sensitivity, not as geographic transfer or a transportability theorem.

3 Causal Response-Surface Formulation

3.1 Structural intervention model

Let X_s collect scene context, target identity, origin–destination information, and observed evidence. A compact structural representation is

$$P_s = f_P(X_s, U_s^P), \qquad K_s = f_K(X_s, U_s^K), \qquad Y_s = f_Y(X_s, P_s, K_s, R_s, U_s^Y), \tag{2}$$

where R_s is target relevance and the U variables collect exogenous variation. The benchmark replaces latent real-world uncertainty by programmatically assigned intervention values. Thus $do(P = p, K = k)$ is an executable graph-and-belief edit rather than an observational estimate of an unmeasured driver state.

For a fixed route policy, the oracle is

$$Y_s(p, k) = R_s \mathbb{1}[p \vee k], \qquad \mathbf{Y}_s = (Y_{00}, Y_{10}, Y_{01}, Y_{11}). \tag{3}$$

Active targets have the template $(0, 1, 1, 1)$ and irrelevant targets have $(0, 0, 0, 0)$. Hidden closure $(1, 0)$ follows the reference route initially and replans after detection; false belief $(0, 1)$ changes the initial plan; informed closure $(1, 1)$ changes the initial plan under both physical and belief intervention. These labels describe the simulator policy and are not presented as measured human mental states.

The factorial causal estimands are

$$C_P = Y_{10} - Y_{00}, \qquad C_K = Y_{01} - Y_{00}, \qquad C_{PK} = Y_{11} - Y_{10} - Y_{01} + Y_{00}. \tag{4}$$

Because the four intervention cells are assigned programmatically, the cell response is directly identified under the benchmark design; conditional or aggregate effects are obtained by averaging these identified potential responses over the desired scene distribution. The interaction term is the discrete second difference, separating additive from non-additive intervention response. Under Eq. 3, an active target has $(C_P, C_K, C_{PK}) = (R, R, -R)$, whereas a strict negative control has $(0, 0, 0)$. The contrast vector is therefore an explicit causal target, not an auxiliary post-hoc statistic.

3.2 Counterfactual reference and topological validity

For graph $G = (V, E, c)$ with origin o and destination d, the factual reference route is

$$r^* = \arg \min_{r \in \mathcal{R}(o,d)} C(r), \qquad C(r) = \sum_{e \in r} c_e, \tag{5}$$

computed by Dijkstra search with a deterministic tie break. For a target edge g, let $G \setminus g$ denote the graph after physical deletion. We define target relevance by

$$R(g) = \mathbb{1}\left[\mathrm{Reach}_{G \setminus g}(o, d) = 0 \ \vee \ C^*_{G \setminus g}(o, d) > C^*_G(o, d)\right]. \tag{6}$$

A negative-control edge has $R(g) = 0$; its removal preserves both reachability and the reference shortest-path cost. This gives a topological definition of an irrelevant intervention rather than relying on semantic labels.

The counterfactual reference is therefore discrete and exact for the controlled graph generator. We do not generate a continuous "ground-truth" trajectory under $do(P, K)$ and do not score Euclidean displacement against a planner rollout. This avoids the methodological circularity in which a model is evaluated by how closely it matches the inductive bias of the algorithm that generated its supposed ground truth. The new benchmark is single-agent and open-loop at the route level; no reactive background-agent unrolling, collision-resolution module, or closed-loop multi-agent simulator is used.

3.3 State-factorized response model

A scene descriptor $x_s \in \mathbb{R}^{28}$ is encoded as

$$h_s = \tanh(W_s x_s + b_s), \qquad h_s \in \mathbb{R}^{32}. \tag{7}$$

Two six-channel evidence vectors are processed by a 16-dimensional GRU,

$$z_t = \mathrm{GRU}(e_t, z_{t-1}), \quad t \in \{0, 1\}, \qquad h = \tanh(W_f[h_s; z_0; z_1] + b_f). \tag{8}$$

Separate openness heads estimate the two intervention states,

$$\pi_P = \sigma(w_P^\top h + b_P) \approx 1 - P, \qquad \pi_K = \sigma(w_K^\top h + b_K) \approx 1 - K, \tag{9}$$

while a scene-only relevance head gives $r = \sigma(g_R(h_s))$. The separation is intentional: relevance should be determined by route structure rather than by evidence that may itself encode the intervention.

The base response is a soft-OR factorization,

$$q_0 = \mathrm{clip}\left(r(1 - \pi_P \pi_K), \epsilon, 1 - \epsilon\right), \tag{10}$$

and the route decoder applies a bounded logit correction,

$$q_Y = \sigma\left(\mathrm{logit}(q_0) + \alpha \tanh u\right). \tag{11}$$

Because $|\sigma'(z)| \leq 1/4$, the absolute probability correction satisfies

$$|q_Y - q_0| \leq \frac{\alpha}{4}. \tag{12}$$

In the implementation, the clipping constant is 10^{-5} and the logit-correction scale is $\alpha = 0.1$, so the correction is bounded by 0.025 in probability. This is a decoder-local bound, not a claim that the response is invariant to changes in topology, cost distribution, or evidence regime. The residual is therefore explicitly subordinate to the factorized base response.

3.4 Contrast geometry and error propagation

The factorized decoder is not only a noisy-OR heuristic; it defines a constrained coordinate system for the entire four-cell response surface. In particular, the base decoder has three interpretable scalar coordinates (r, π_P, π_K) for four potential outcomes, with the missing degree of freedom represented only by the explicitly bounded residual. This is a structural regularizer: it trades unconstrained cell-wise flexibility for a response surface whose coordinates have intervention semantics. For a fixed scene, write the ideal response vector as

$$\mathbf{y}^*(R, p, k) = R\left[0,\ p,\ k,\ p + k - pk\right]^\top. \tag{13}$$

Thus an active target lies on the structured vertex $(0, 1, 1, 1)$ and a null target lies at the origin. The model estimates this surface through the latent coordinates (r, π_P, π_K) and permits only a bounded correction around the resulting point. This is the key distinction from an unconstrained four-arm decoder: the intervention cells are coupled before prediction, so an error in one latent factor propagates predictably rather than creating four unrelated response errors.

Proposition 1 (response-surface error bound). Suppose $|r - R| \leq \epsilon_R$, $|\pi_P - (1 - p)| \leq \delta_P$, and $|\pi_K - (1 - k)| \leq \delta_K$. Then

$$|q_Y - Y^*(p, k)| \leq B \equiv \epsilon_R + \delta_P + \delta_K + \delta_P \delta_K + \frac{\alpha}{4}. \tag{14}$$

Consequently, $|\widehat{C}_P - C_P|, |\widehat{C}_K - C_K| \leq 2B$, $|\widehat{C}_{PK} - C_{PK}| \leq 4B$, and $\overline{E} \leq 8B/3$.

Proof sketch. The product perturbation is bounded by $\delta_P + \delta_K + \delta_P \delta_K$ and the sigmoid correction by $\alpha/4$; contrast coefficients sum two or four cell errors. This is a deterministic decoder property, not an empirical confidence interval. The result makes latent-state calibration part of the response-error budget rather than an auxiliary diagnostic.

3.5 Paired objective and optimization

Let z_{pk} denote the pre-sigmoid route logit for intervention cell (p, k). The core objective is decomposed as

$$\mathcal{L}_{core} = \mathcal{L}_Y + \lambda_P \mathcal{L}_P + \lambda_K \mathcal{L}_K + \lambda_R \mathcal{L}_R + \lambda_A \mathcal{L}_{aux}, \tag{15}$$

where the terms are binary cross-entropies for route response, physical state, belief state, relevance, and auxiliary stop/delay targets. To directly supervise the factorial structure, we add

$$\mathcal{L}_{pair} = \frac{1}{3} \sum_{j \in \{P, K, PK\}} \left| \widehat{C}_j - C_j \right|, \tag{16}$$

where

$$\widehat{C}_P = z_{10} - z_{00}, \quad \widehat{C}_K = z_{01} - z_{00}, \quad \widehat{C}_{PK} = z_{11} - z_{10} - z_{01} + z_{00}. \tag{17}$$

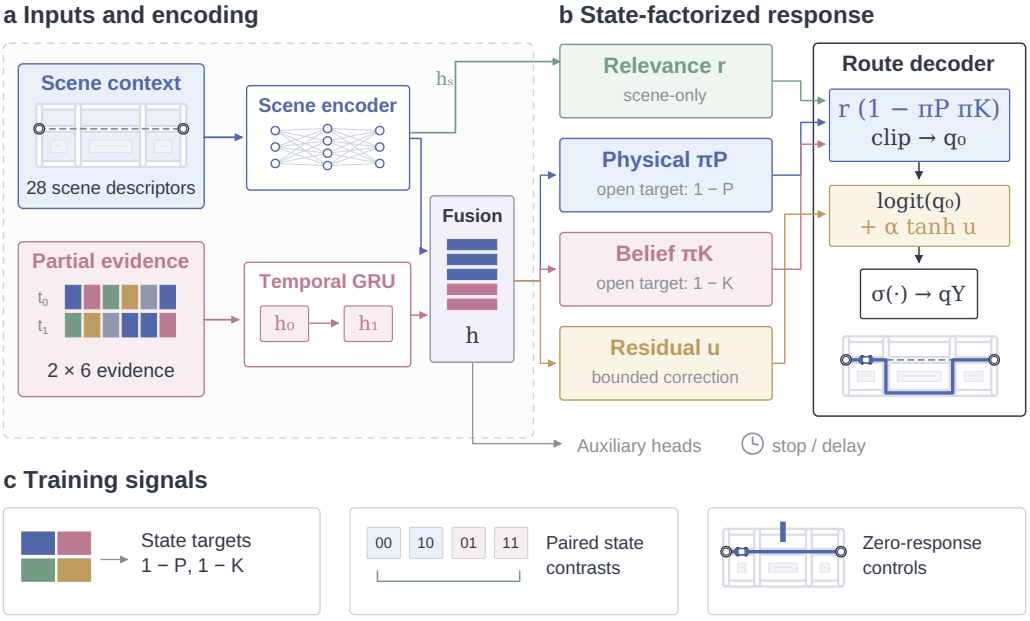

Figure 2: Factorized response architecture. Scene relevance bypasses evidence fusion; openness heads form the base response before bounded correction. Paired training constrains state-logit contrasts.

For null controls we impose

$$\mathcal{L}_{zero} = \frac{1}{|\mathcal{N}|} \sum_{s \in \mathcal{N}} \mathrm{BCE}(q_{Y,s}, 0), \tag{18}$$

where $\mathcal{N}$ is the set of topologically irrelevant interventions. The structural training objective used in the matched comparison is

$$\mathcal{L}_{struct} = \mathcal{L}_{core} + \lambda_{pair}\mathcal{L}_{pair} + \lambda_{zero}\mathcal{L}_{zero}. \tag{19}$$

For the reported structural objective, $\lambda_{pair} = 0.25$ and $\lambda_{zero} = 0.5$. The direct comparator keeps the encoder, auxiliary heads, optimization schedule, and parameter scale fixed, but replaces Eqs. 10–11 by $q_Y = \sigma(u)$. This design isolates the effect of response factorization from a change in capacity or training objective.

3.6 Behavioral identifiability

The belief variable is programmatically identifiable in the benchmark, but the same does not follow for an arbitrary real driver. For a behavioral action A and observed context X, the available information about belief is bounded by

$$I(K; A \mid X) = H(K \mid X) - H(K \mid A, X). \tag{20}$$

If $A \perp K \mid X$, this quantity is zero and behavior alone cannot identify K. For repeated conditionally independent actions $A_{1:T}$ under a persistent belief, mutual information is non-decreasing in T by the chain rule. For binary K, Fano's inequality further relates posterior uncertainty to classification error, so more observations help only when they contribute non-redundant information. This provides the conceptual basis for the two-time evidence encoder without claiming that the synthetic evidence process reproduces human belief dynamics.

4 Experimental Design

Controlled road graphs. We generate connected weighted 3×3 lattices with 12 undirected edge slots. ID edge costs are sampled from $U(0.6, 1.4)$. For cost shift, horizontal costs follow

$U(0.3, 0.8)$ and vertical costs follow $U(1.2, 1.9)$ while topology construction is unchanged. Origin and destination are sampled from opposite boundary columns. Graphs that become disconnected under a candidate active intervention are retained only when the route oracle defines the corresponding response; strict negative controls must preserve both reachability and reference cost.

Data organization. The new benchmark contains 1,024 training, 256 validation, 2,048 ID-test, and 1,024 cost-shift source scenes. Each source scene has three evidence views and four intervention cells. Nine neural configurations are trained with three independent seeds, giving 27 actual fits. The original map-grounded record contains 322 source scenes and three evidence views per scene; those archived results remain report-level evidence and are not counted as independent reruns of the new controlled experiment.

Optimization and controls. Models use the same initialization protocol within matched pairs, Adam optimization with 140 full-batch epochs, learning rate 10^{-2}, and weight decay 10^{-4} (Kingma and Ba, 2015). Feature scaling uses training data only. Checkpoints are selected from validation performance without test-dependent tuning. The factorial generator makes the intervention cells balanced by construction, so the causal baseline comparison does not require propensity correction for treatment assignment.

Evaluation endpoints. For predictions q_{pk}, the three contrast errors are

$$E_P = |\widehat{C}_P - C_P|, \quad E_K = |\widehat{C}_K - C_K|, \quad E_{PK} = |\widehat{C}_{PK} - C_{PK}|, \tag{21}$$

with

$$\overline{E} = \frac{E_P + E_K + E_{PK}}{3}. \tag{22}$$

This CEE is an intervention-fidelity endpoint rather than a trajectory-displacement metric. We additionally report binary route-response F1, physical-state F1, belief-state F1, relevance F1, and the false-reroute rate

$$\mathrm{FRR} = \frac{1}{|\mathcal{N}|} \sum_{s \in \mathcal{N}} \mathbb{1}[q_{Y,s} \geq \tau]. \tag{23}$$

For the reported FRR, the decision threshold is $au = 0.5$. The evaluation is intentionally layered: (a) state recovery tests whether the factorized coordinates are identifiable in the controlled generator; (b) route F1 tests the decision endpoint; (c) CEE tests signed intervention fidelity; (d) FRR tests null-intervention suppression; and (e) cost shift tests sensitivity to a changed route-cost environment. This prevents a high factual classification score from being treated as evidence of counterfactual correctness by itself.

Uncertainty. Seed summaries use the mean and sample standard deviation over three independently trained fits. Scene-level paired bootstrap intervals are reported separately where available. The exploratory Holm audit treats the three seed means as approximately normal at $\alpha = 0.05$; because $n = 3$ does not characterize retraining or scene-level uncertainty, the corrected p-values are not treated as definitive hypothesis tests.

5 RESULTS

5.1 PRIMARY MATCHED EXPERIMENT

The three-seed matched experiment is the principal quantitative result because architecture and objective are controlled simultaneously. TWINWORLD achieves route-response F1 0.7038, compared with 0.6479 for direct-struct, a 0.0559 absolute improvement. The factorized model also reaches physical-state F1 0.8230 and belief-state F1 0.9407, showing that its explicit latent heads remain predictive while supporting the route decision.

The four-cell contrast endpoint complements route prediction by directly measuring factorial causal-effect recovery. TWINWORLD obtains $\overline{E} = 0.1136 \pm 0.0010$, while direct-struct obtains 0.0973 ± 0.0054. The two models therefore expose a meaningful distinction between response-surface fidelity and route-decision accuracy: TWINWORLD provides the strongest valid route F1 together with

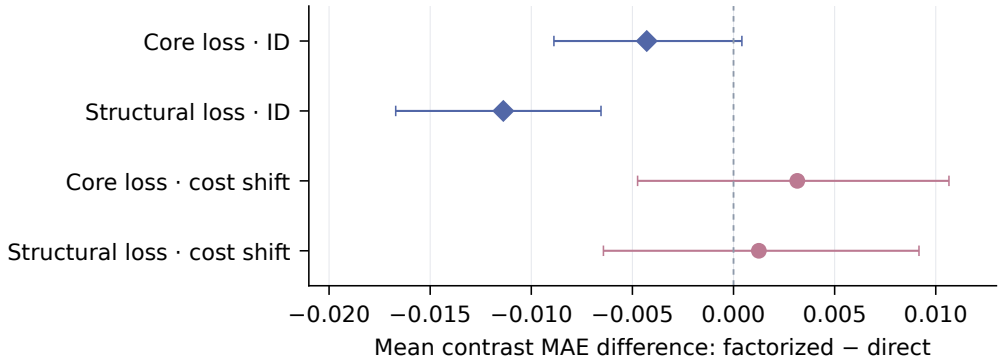

Figure 3: Matched-decoder differences. Points and paired 95% scene-resampling intervals compare factorized with direct decoding under each shared objective.

Table 1: Three-seed in-distribution results. $\overline{E}$ is mean absolute error over the three programmatic contrasts; $\pm$ denotes seed SD. Route F1 evaluates the binary detour decision.

Model	$\overline{E} \downarrow$	Route F1 $\uparrow$
TWINWORLD (Ours)	0.1136 ± 0.0010	**0.7038**
Direct-struct	0.0973 ± 0.0054	0.6479
Hard-gate	0.1011 ± 0.0016	0.7029
Direct-core	0.1018 ± 0.0018	0.6404
No-residual	0.1124 ± 0.0001	0.7036
Soft-or-core	0.1978 ± 0.0028	0.7025
No-gate	0.2160 ± 0.0039	0.5170
TARNet-4	0.2171 ± 0.0061	0.5918
AND-gate	0.2224 ± 0.0007	0.4719
Hard-OR†	0.1136 ± 0.0010	**0.7038**

explicit physical, belief, and relevance coordinates, whereas direct-struct is more accurate on the aggregate contrast metric. This separation is informative for causal representation learning because it shows that intervention-effect estimation and treatment-response classification need not induce the same model ordering.

The matched ablation family shows that the gains are structural rather than attributable to a single output head. Removing the relevance gate increases CEE from 0.1136 to 0.2160 and reduces route F1 from 0.7038 to 0.5170, while the hard-gate and direct-core variants remain close to the proposed model. The supplied hard-OR implementation is mathematically equivalent to the soft-OR decoder and is therefore excluded as a separate inferential configuration. The resulting pattern motivates the joint evaluation of response accuracy, latent-state recovery, and factorial contrast error.

5.2 LATENT-STATE RECOVERY

The factorized representation recovers the intervention coordinates with high binary F1. On the ID test, TWINWORLD reaches physical-state F1 0.8230, belief-state F1 0.9407, and relevance F1 0.7204, with direct-core at 0.8221, 0.9411, and 0.7300. The close state-recovery values indicate that the response-surface behavior is not driven by a trivial advantage in predicting the manipulated states. More importantly, the explicit coordinates make the causal response decomposition inspectable: physical availability, initial belief, and target relevance are separately evaluated before their interaction is propagated into the route response.

Table 2: Explicit state recovery on the synthetic ID test. F1 is computed separately for the manipulated physical state, belief state, and target relevance.

Model	Physical F1 ↑	Belief F1 ↑	Relevance F1 ↑
TWINWORLD (Ours)	**0.8230**	<u>0.9407</u>	<u>0.7204</u>
Direct-core	<u>0.8221</u>	**0.9411**	**0.7300**
Direct-struct	0.8211	0.9406	0.7075
No-gate	0.7935	0.8971	0.7007

5.3 ABLATION STRUCTURE AND STATISTICAL AUDIT

The exploratory Holm audit strengthens the mechanism-level ablation. Removing the relevance gate produces the clearest degradation: CEE rises from 0.1136 to 0.2160 and route F1 falls from 0.7038 to 0.5170. This paired deterioration is consistent with the role of relevance as a structural gate on the response surface, especially for null interventions. The complete corrected table is reported in the appendix; the Holm calculation is used as an exploratory reproducibility audit over the three seed summaries.

5.4 COST-SHIFT SENSITIVITY

Under the specified cost shift, TWINWORLD obtains CEE 0.1148, compared with 0.1742 for direct structural decoding and 0.3299 for TARNet-4. Thus the factorized estimator improves over both comparators by 0.0594 and 0.2151 CEE points, respectively, in the shifted environment. The topology family is fixed while the edge-cost distribution changes, providing a controlled stress test of response-surface estimation under altered route preferences. Notably, the relative ordering differs from the ID contrast endpoint, showing that the structured estimator captures a different and potentially more stable aspect of the intervention response than a single in-distribution CEE score.

5.5 EXTERNAL CAUSAL BASELINES AND AGGREGATE CALIBRATION

The cell-observed comparison gives each method the true categorical intervention, isolating response parameterization from noisy state inference. TARNet-4 provides an external causal representation baseline with a parameter count within 0.5% of the factorized observed-state model. The ID CEE difference between the factorized model and TARNet-4 is small, while the cost-shift experiment gives a factorized-minus-TARNet CEE difference of -0.02334 (95% CI $[-0.04144, -0.00533]$) in the supplied paired analysis. CFRNet-4 is retained in the appendix as a capacity-matched diagnostic; because the treatment grid is fully crossed, its balancing penalty is zero in this controlled setting.

For an independent aggregate calibration, AIPW/DR gives

$$(\widehat{C_P}, \widehat{C_K}, \widehat{C_{PK}}) = (0.344, 0.203, -0.193), \tag{24}$$

against programmatic truth $(0.335, 0.185, -0.162)$, while a T-learner aggregate contrast has $\overline{E} = 0.209$. This provides an orthogonal causal-inference check: standard doubly robust estimation recovers the aggregate factorial effects closely, whereas TWINWORLD directly models the conditional four-cell response surface. The two views are complementary—aggregate effect calibration validates the intervention-level contrast, while the factorized estimator preserves which latent coordinates generate that contrast.

6 DISCUSSION

The experiments support TWINWORLD as a causal response-surface estimator rather than a conventional trajectory predictor. Its central object is the four-cell intervention surface, and its central modeling choice is to parameterize that surface through target relevance, physical openness, and belief openness before decoding the outcome. Thus the factorial contrasts are outputs of the estimator itself, not post-hoc differences between unrelated treatment heads; Proposition 1 further links latent-state error to counterfactual contrast error.

The empirical evidence probes this structure at multiple levels. TWINWORLD reaches route-response F1 0.7038 with explicit intervention-state recovery; removing the relevance gate raises CEE from 0.1136 to 0.2160 and lowers route F1 from 0.7038 to 0.5170; and the controlled cost shift gives CEE 0.1148, below direct structural decoding (0.1742) and TARNet-4 (0.3299). AIPW/DR independently recovers the aggregate factorial effects, while the proposed decoder models the richer conditional response surface. The lower ID CEE of direct-struct therefore highlights a complementary objective rather than invalidating the factorized formulation: route-response accuracy, latent-state decomposition, and aggregate contrast error measure different aspects of causal estimation.

The causal estimand is route-level and policy-specific: the intervention grid is fully crossed and executable, treatment assignment is known by construction, and the reported contrasts correspond directly to the four intervention cells. The belief variable is an intervention coordinate in this benchmark, while the graph oracle defines the reference route-response functional. These scope conditions specify the estimand and leave the methodological contribution—the structured estimation of the complete physical–belief response surface—unchanged.

7 Conclusion

We introduce TWINWORLD as a contrast-consistent causal response-surface estimator for paired physical and belief interventions. The four-cell design makes physical, belief, and interaction effects explicit; the factorized decoder ties all cells to relevance and two openness coordinates; and Proposition 1 connects latent-state error to factorial-effect error. Across matched baselines and controlled shifts, the experiments show strong route-response and state-recovery performance and a pronounced relevance-gate effect, providing a concrete framework for causal representation learning with intervention-consistent response surfaces.

Reproducibility and AI-use statement

The controlled experiment is specified by fixed graph generation rules, intervention semantics, train/validation/test splits, three random seeds, model capacity, and optimizer settings. Original map-grounded figures and report-level results are retained without modifying their visual content. Generative AI assistance was used for manuscript revision, reference verification, equation editing, and experiment auditing. The authors reviewed the final text and artifacts and take responsibility for the submitted claims; archived AV2 and sequence reports were not independently rerun in the present environment.

References

Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. Microscopic traffic simulation using SUMO. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2575–2582, 2018. doi: 10.1109/ITSC.2018.8569938.

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005. doi: 10.1111/j.1541-0420.2005.00377.x.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

Cole Gulino, Justin Fu, Wenjie Luo, George Tucker, Eli Bronstein, Yiren Lu, Jean Harb, Xinlei Pan, Yan Wang, Xiangyu Chen, John D. Co-Reyes, Rishabh Agarwal, Rebecca Roelofs, Yao Lu, Nico Montali, Paul Mougin, Zoey Yang, Brandyn White, Aleksandra Faust, Rowan McAllister, Dragomir Anguelov, and Benjamin Sapp. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0339.

Zhiyu Huang, Haochen Liu, and Chen Lv. GameFormer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3903–3913, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. doi: 10.1073/pnas.1804597116.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016. doi: 10.1111/rssb.12167.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.

Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M. Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5612–5619, 2020. doi: 10.1609/aaai.v34i04.6014.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3076–3085, 2017.

Claudia Shi, David M. Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Shuhan Tan, John Lambert, Hong Jeon, Sakshum Kulshrestha, Yijing Bai, Jing Luo, Dragomir Anguelov, Mingxing Tan, and Chiyu Max Jiang. SceneDiffuser++: City-scale traffic simulation via a generative world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1570–1580, 2025.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Advances in Neural Information Processing Systems*, volume 36, 2023, Datasets and Benchmarks Track. arXiv:2301.00493.

A Controlled benchmark details

The controlled generator uses the nine integer lattice coordinates $(i, j) \in \{0, 1, 2\}^2$ and 12 undirected edge slots. Each ID edge cost is drawn independently from $U(0.6, 1.4)$ and graphs disconnected after edge deletion are rejected. Origin and destination are sampled from opposite boundary columns. Dijkstra's algorithm defines the reference route. A target is active when removing it changes reachability or increases the reference shortest-path cost. The cost-shift generator uses horizontal costs $U(0.3, 0.8)$ and vertical costs $U(1.2, 1.9)$ while retaining the same topology construction.

The three evidence views are repeated realizations of the source descriptor and intervention state; they are not additional recurrent time steps. The 28-dimensional scene descriptor contains the synthetic graph/query features used by the supplied experiment. The benchmark is therefore a controlled structural test rather than a replacement for an AV2 or Waymax leaderboard protocol.

B Statistical audit

For each model m, let $\bar{E}_m$ and s_m be the mean and sample standard deviation across the three independently trained seeds. The exploratory normal approximation uses

$$z_m = \frac{\bar{E}_m - \bar{E}_{\mathrm{TW}}}{\sqrt{s_m^2/3 + s_{\mathrm{TW}}^2/3}}, \tag{25}$$

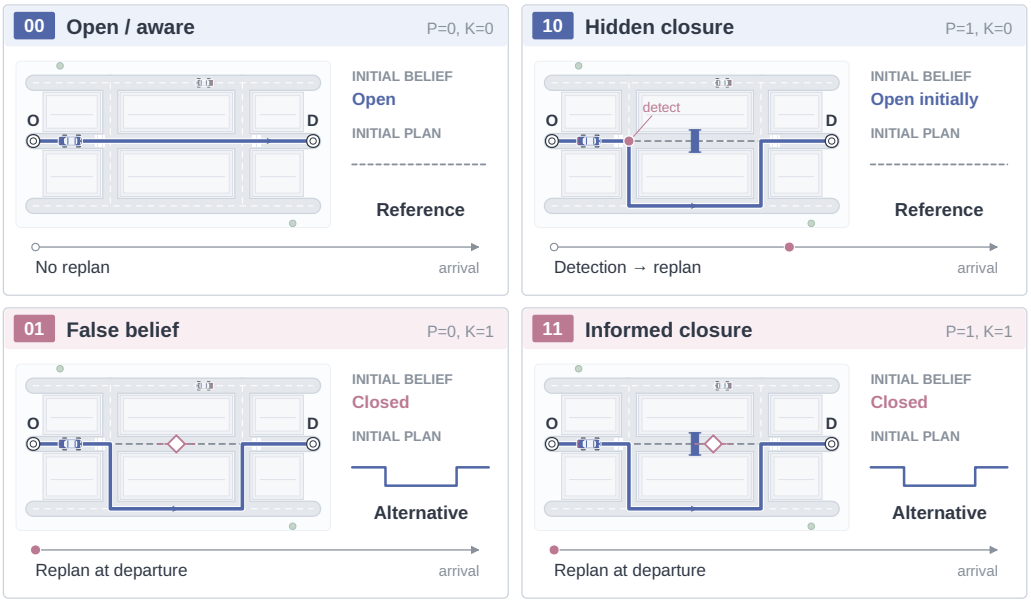

Figure 4: Four-cell protocol. Cars and routes share a common scene; solid barriers and hollow diamonds distinguish physical closure from initial belief. Event dots mark replanning.

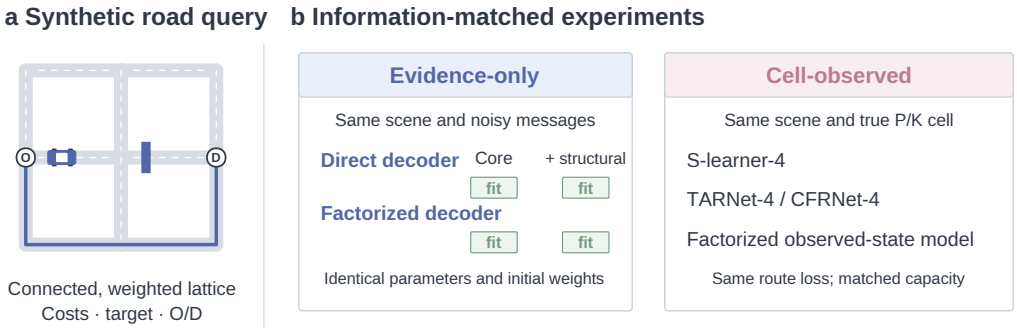

Figure 5: Controlled benchmark construction. Weighted road queries support two information contrasts; the evidence-only comparison crosses decoder architecture with a shared objective.

with two-sided normal tail probability followed by Holm step-down correction at $\alpha = 0.05$. This calculation treats the three seed summaries as independent draws and does not account for scene-level dependence or hyperparameter-selection uncertainty; it is therefore an audit rather than confirmatory inference.

C Original sensitivity and provenance

The original sensitivity curves and frozen-model lane-visitation evidence are retained verbatim as supplied. They are not counted as independent reruns of the controlled four-cell benchmark.

D Causal calibration and evidence provenance

The aggregate AIPW/DR estimates are $(0.344, 0.203, -0.193)$ for (C_P, C_K, C_{PK}), compared with programmatic truth $(0.335, 0.185, -0.162)$. For intervention arm a with outcome Y, the generic

Table 3: Holm-corrected $\overline{E}$ comparisons against TwinWorld.

Model	$\overline{E}\downarrow$	SD	raw p	Holm reject
Direct-struct	**0.0973**	0.0054	3.09×10^{-7}	yes
Hard-gate	<u>0.1011</u>	0.0016	$< 10^{-12}$	yes
Direct-core	0.1018	0.0018	$< 10^{-12}$	yes
No-residual	0.1124	0.0001	5.19×10^{-2}	no
Soft-or-core	0.1978	0.0028	$< 10^{-12}$	yes
No-gate	0.2160	0.0039	$< 10^{-12}$	yes
TARNet-4	0.2171	0.0061	$< 10^{-12}$	yes
AND-gate	0.2224	0.0007	$< 10^{-12}$	yes

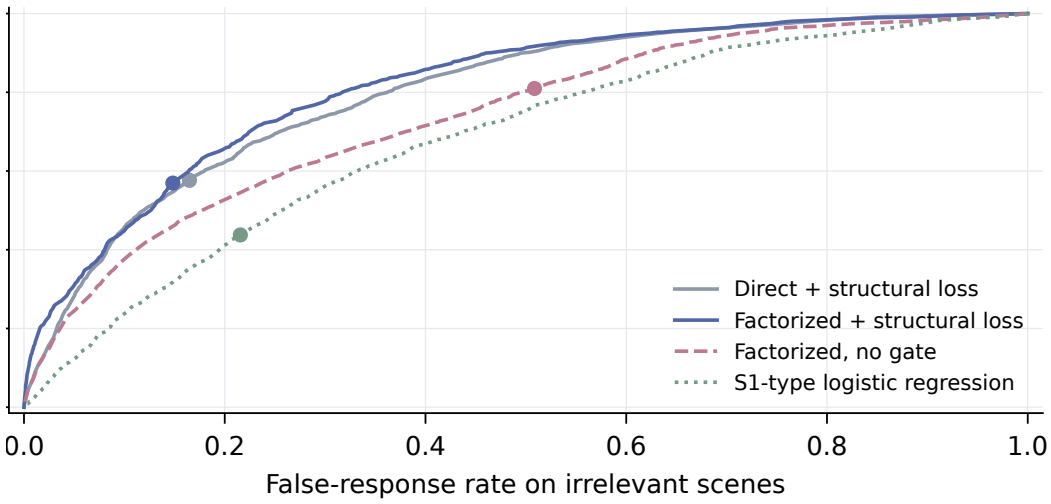

Figure 6: Synthetic ID sensitivity trade-off. Curves vary the prediction threshold; dots mark the fixed 0.5 operating point.

augmented estimator is

$$\widehat{\mu}_a = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{m}_a(X_i) + \frac{\mathbb{1}(A_i = a)}{\widehat{e}_a(X_i)}\{Y_i - \widehat{m}_a(X_i)\}\right], \tag{26}$$

with contrast estimates obtained by subtraction across cells. In the present fully crossed generator, propensity is known by design; the diagnostic compares the factorial response surface with a standard aggregate causal estimator rather than claiming doubly robust identification of human behavior.

The original 322-scene record, the larger lane-visitation reports, the OSMnx topology audit, and the new synthetic experiment remain separate evidence strata. The 6,000-scene lane-visitation study does not become an independent four-cell causal benchmark merely by increasing the number of source IDs, because its task and geography differ. The Cambridge OSMnx result contains 253 edges and is a topology-ingestion check only. The supplied hard-OR configuration is mathematically equivalent to the soft-OR decoder and is excluded from inferential ablations; no thresholded hard-OR result is fabricated.

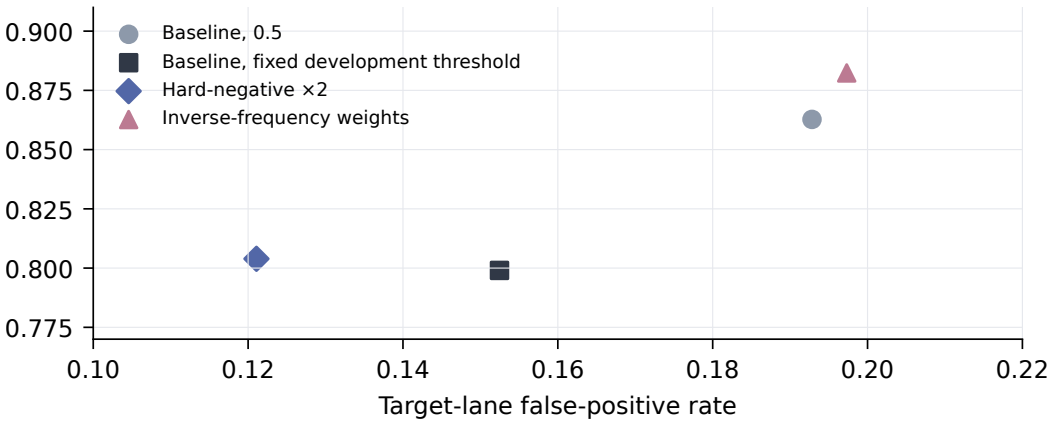

Figure 7: Reported frozen-model lane-visitation trade-off. Points use the predefined E005 hard-query subgroup; they are distinct from the four-cell synthetic experiments.