CONSTRAINING SHORTCUT PATHWAYS: A FRONT-DOOR-INSPIRED INTERACTION BOTTLENECK FOR END-TO-END AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

End-to-end autonomous driving exchanges information among traffic participants, map elements, and ego planning states through high-dimensional latent representations. Although expressive, these representations can also encode contextual cues that are predictive during training but weakly tied to physical interaction. When propagated across entities, such cues can form shortcut pathways that make planning sensitive to distribution-specific context. To constrain this information flow, we introduce the **Causal Interaction Bottleneck** (CIB), a front-door-inspired module for cross-entity interaction. For each source, CIB predicts type-specific kinematic or geometric hypotheses and associated mode probabilities. Each hypothesis is re-encoded before CIB performs target-conditioned attention, with the mode probabilities serving as priors. Raw source features do not bypass this mediated path. At the network level, this design enforces source-side structural mediation: newly transmitted source information can affect the target only through the mediator. Under deterministic mediation, the mediator cannot increase contextual information after conditioning on the physical anchor and entity type. We further establish stability of the resulting attention rule and derive explicit bounds for the discrete kinematic update. Across sequential, parallel, and iterative architectures, CIB consistently improves planning performance. Matched controls indicate that these gains are not fully explained by generic bottleneck regularization. Representation probes and paired interventions further show that CIB reduces sensitivity to appearance-only changes while preserving responses to behaviorally relevant physical changes.

1 INTRODUCTION

End-to-end autonomous driving has made substantial progress by jointly modeling perception, motion prediction, and planning within a unified architecture (Hu et al., 2023; Jiang et al., 2023; Sun et al., 2025). This integration reduces the need for hand-designed interfaces but also makes planning directly dependent on high-dimensional latent representations. These representations can encode task-relevant interaction cues alongside visual appearance, scene context, and statistical regularities specific to the training distribution. When such regularities shift at test time, the planner may continue to rely on correlations that were predictive during training but are not stably grounded in the underlying physical interaction (de Haan et al., 2019; Geirhos et al., 2020).

This vulnerability also depends on how learned representations are routed through the architecture. Sequential designs can pass prediction-stage representations to the planning module beyond explicit motion states (Hu et al., 2023; Jiang et al., 2023). Parallel designs jointly model motion prediction and ego planning through high-dimensional query interactions (Sun et al., 2025; 2026). Iterative designs repeatedly update and reuse interaction states during refinement (Chen et al., 2024). Although their implementations differ, these architectures create distinct routes through which contextual information can enter or be reused by planning. We refer to these computation-level routes as *architectural shortcut pathways*. The term describes information flow permitted by the network rather than a causal graph of the physical world. This view motivates constraining what information can cross selected interaction edges instead of enumerating every possible contextual factor.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

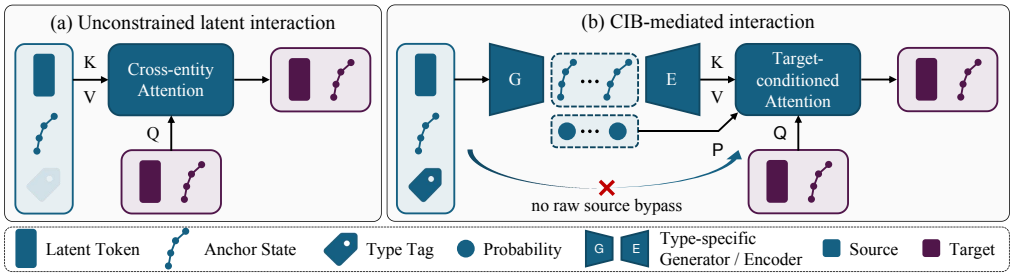


Figure 1: **Architectural shortcut pathways and physical mediation.** Conventional interaction modules allow source-specific latent features and anchor-derived representations to enter cross-entity communication directly. CIB instead routes source-dependent interaction content through a typed mediator representation consisting of physical hypotheses and their mode probabilities.

Motivated by the mediated influence structure emphasized by the front-door criterion (Pearl, 2009), we introduce the **Causal Interaction Bottleneck** (CIB). As illustrated in Figure 1, CIB first predicts type-specific kinematic or geometric hypotheses and their mode probabilities for selected source-to-target interactions. Each hypothesis is independently re-encoded and then participates in target-conditioned attention, where the mode probabilities act as priors. Raw source features are not reintroduced on the replaced interaction edge. Physical states therefore serve as required mediator representations for source information transmitted through selected cross-entity interactions, rather than merely as prediction outputs or auxiliary constraints.

The front-door criterion provides structural motivation for this design, not an observational causal-effect estimator. We establish an architecture-level *structural mediation* property: once the target state, active source set, attention mask, and remaining source mediators are fixed, two source descriptors that induce the same mediator representation produce the same CIB update. Deterministic mediation further yields a conditional information non-increase result. We also analyze the stability of target-conditioned attention to perturbations of the mediator distribution and derive explicit physical bounds for the discrete kinematic update. These results characterize the network computation without assuming the graphical or statistical conditions for classical front-door identification.

We evaluate CIB at both the performance and mechanism levels. Experiments span sequential, parallel, and iterative architectures under open-loop, non-reactive, and reactive closed-loop protocols. A matched Direct-State Bottleneck separates generic bottleneck regularization from explicit physical integration. Frozen representation probes and paired simulator interventions examine whether reduced appearance sensitivity is accompanied by preserved responses to behaviorally relevant physical changes. Same-backbone comparison with a complementary intervention method and transfer to a stronger planner further test the scope and portability of the proposed communication constraint.

Our primary contributions are summarized as follows: (i) We provide a unified information-flow view of architectural shortcut pathways in end-to-end driving and formulate the problem as constraining information transmitted across selected interaction edges. (ii) We introduce CIB, which uses explicit physical states as mediator representations in cross-entity interaction, and characterize its structural mediation, information non-increase, attention stability, and discrete physical bounds. (iii) We connect these structural properties to planning performance and robustness through matched controls, representation probes, paired interventions, and cross-architecture evaluation.

2 Related Work

End-to-end driving architectures. Modern end-to-end driving has evolved from direct sensor-to-control prediction toward unified perception, prediction, and planning systems. UniAD organizes multiple tasks through planning-oriented query interfaces (Hu et al., 2023), VAD represents agents and map elements as vectorized planning constraints (Jiang et al., 2023), and SparseDrive jointly models perception, motion prediction, and planning with sparse scene representations (Sun et al., 2025). PPAD further models dynamic feedback through iterative prediction-planning interaction (Chen et al., 2024). Recent systems continue to improve planner capacity and closed-loop per-

formance from different directions: ORION combines vision-language reasoning with generative planning (Fu et al., 2025), Hydra-NeXt uses control prediction and kinematics-aware refinement (Li et al., 2025), and SparseDriveV2 expands action-space coverage through factorized path–velocity representations and coarse-to-fine scoring (Sun et al., 2026). Whereas these methods primarily improve planning representations, reasoning, or trajectory-space coverage, CIB focuses on what source-side information is allowed to enter a cross-entity interaction.

Causal robustness in driving. Causal confusion shows that imitation policies can exploit observations that are highly predictive of expert actions yet fail under distribution shift (de Haan et al., 2019). CausalVAD brings this concern to modern end-to-end driving and introduces the Sparse Causal Intervention Scheme, which uses learned prototype dictionaries for backdoor-inspired latent intervention (Tang et al., 2026). CIB acts at a different point of the representation pipeline. Rather than constructing a dictionary of contextual factors, it constrains cross-entity communication through physically grounded mediators. We therefore use the front-door criterion as structural motivation for mediated information flow rather than treating CIB as a neural estimator of the classical front-door adjustment. More broadly, causal representation learning shows that latent causal identification generally requires additional assumptions on environments, interventions, supervision, or the data-generating process (Schölkopf et al., 2021; Ng et al., 2025).

Physics-grounded prediction and planning. Physical structure has long been used to improve the feasibility of learned prediction and planning. Neural motion planners have combined learned representations with physically feasible trajectory candidates (Zeng et al., 2019); subsequent work has incorporated kinematic bicycle models into trajectory prediction (Liao et al., 2024), while systems such as Hydra-NeXt employ kinematics-aware refinement for closed-loop execution (Li et al., 2025). The distinction of CIB is therefore not the use of vehicle kinematics or vectorized geometry itself, but their role in the network. The resulting physical states are not merely prediction outputs, trajectory candidates, or auxiliary constraints; they serve as mediator representations on replaced cross-entity interaction edges and thereby directly constrain source-specific interaction content.

3 METHODOLOGY

End-to-end driving models typically exchange information among traffic participants, map elements, and ego planning states through high-dimensional latent representations. Although expressive, such representations also allow contextual information unrelated to the underlying physical interaction to enter cross-entity communication directly. Our objective is not to identify a real-world causal effect from observational driving data. Instead, we impose an explicit structural constraint on source-to-target information flow inside the network. We introduce the **Causal Interaction Bottleneck** (CIB). For each replaced interaction edge, CIB maps the source descriptor to a set of physically grounded motion or geometric hypotheses and their predicted probabilities. Each hypothesis is independently encoded as an interaction token and remains distinct until target-conditioned attention. The raw source representation is not used directly as a source-side key or value. Figure 2 summarizes the complete CIB forward computation and its instantiation across the three interaction paradigms.

3.1 ARCHITECTURAL SHORTCUT PATHWAYS AND DESIGN OBJECTIVE

For source entity s, let $q_s \in \mathbb{R}^d$ denote its latent query, b_s the physical anchor used to initialize the corresponding integrator, and τ_s its entity type. We define the source descriptor as $\chi_s = (q_s, b_s, \tau_s)$ and denote the target latent query by q_t.

Conventional interaction modules can construct source-dependent messages directly from latent features and anchor-derived representations. Appearance, scene context, and historical statistics encoded in these representations can therefore influence the target update. CIB instead maps χ_s to a mediator representation $\mathcal{M}_s$. Conditional on the active source set and attention mask, all newly transmitted source-specific information on a CIB-replaced edge must pass through $\mathcal{M}_s$, while the target retains its own residual state.

Sequential, parallel, and iterative architectures can expose different information-flow paths through prediction-stage context, bidirectional latent message passing, and recurrent reuse of latent states, respectively. We refer to these computation-level routes as *architectural shortcut pathways*. Our

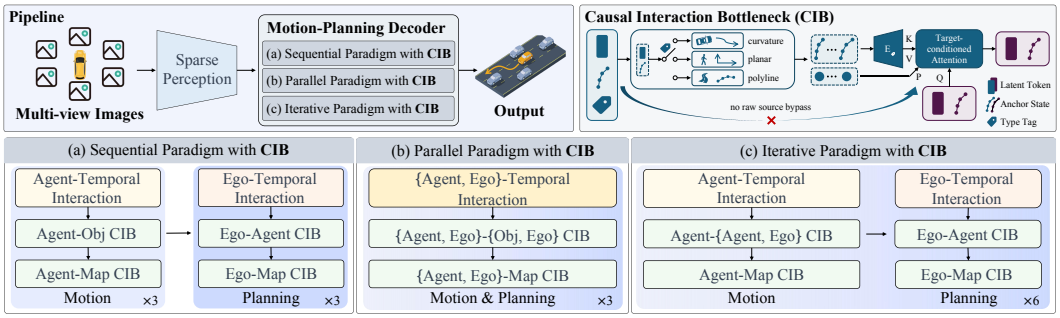


Figure 2: **CIB forward computation and cross-paradigm instantiation.** Given a source descriptor, CIB predicts mode-specific physical parameters and probabilities. Entity type selects curvature-based, planar-motion, or intrinsic-polyline integration, and every resulting physical hypothesis is independently re-encoded as a source-mode token. Predicted mode probabilities enter target-conditioned attention as log priors, and all valid source-mode tokens are normalized jointly. Raw source features do not bypass the mediator on replaced interaction edges. The bottom panels show sequential, parallel, and iterative instantiations.

design objective is to constrain newly transmitted source information on selected interaction edges rather than to infer a causal graph of the physical world from observational driving data.

3.2 Front-Door-Inspired Structural Mediation

The classical front-door criterion shows that, under specific graphical conditions, the effect of a treatment on an outcome can be identified from observational data through a mediator (Pearl, 2009). We borrow only the structural principle that source influence should pass through an explicit mediator, without assuming that learned source representations, physical mediators, and target states satisfy the graphical and statistical conditions required for classical identification.

For source s, CIB maps the source descriptor $\chi_s = (q_s, b_s, \tau_s)$ to

$$\mathcal{M}_s = \mathcal{G}(\chi_s) = \left(\tau_s, \left\{N_s^{(k)}, p_s^{(k)}\right\}_{k=1}^{K_s}\right), \quad \sum_{k=1}^{K_s} p_s^{(k)} = 1, \tag{1}$$

where $N_s^{(k)}$ is the k-th physical hypothesis and $p_s^{(k)}$ is its predicted probability. The defining constraint of a CIB-replaced interaction is that, once the active source set and attention mask are fixed, newly transmitted source-specific information can reach the target only through $\mathcal{M}_s$.

Proposition 1 (Structural mediation)**.** *Fix the target query, active source set and attention mask, and the mediator representations of all sources other than s. If two source descriptors satisfy $\mathcal{G}(\chi_s) = \mathcal{G}(\widetilde{\chi}_s)$, then the corresponding CIB interaction produces exactly the same target message and target update under these fixed conditioning variables.*

This property follows from the computation graph and is the architecture-level claim associated with CIB. To characterize its information-theoretic consequence, let Z denote any contextual variable defined on the same probability space and let $B_s = (b_s, \tau_s)$. Since $\mathcal{M}_s$ is a deterministic function of (q_s, B_s), the conditional data processing inequality gives

$$I(Z; \mathcal{M}_s \mid B_s) \le I(Z; q_s \mid B_s) \tag{2}$$

(Cover & Thomas, 2006). Equation 2 states that deterministic mediation cannot increase the information about Z available through the mediated source representation. It neither requires a strict reduction nor implies independence between Z and $\mathcal{M}_s$.

The analysis above treats $\mathcal{G}$ as an abstract mediator map. We next instantiate this map with the differentiable physical generator and prior-weighted interaction used by CIB. Formal statements and proofs are provided in Appendix A.

3.3 CAUSAL INTERACTION BOTTLENECK

Section 3.2 defines $\mathcal{G}$ abstractly as a map from a source descriptor to a mediator representation. We now instantiate this map with a differentiable physical generator followed by prior-weighted interaction. CIB first constructs a finite distribution over physical hypotheses and then uses only these hypotheses and their probabilities to transmit newly generated source-specific information to the target on each replaced interaction edge.

Physical-hypothesis generation. The source query is processed by a shared generator, $h_s = \Phi_G(q_s)$. For dynamic entities, mode k uses a learned mode embedding $e^{(k)}$. A mode-scoring head and a type-specific physical head produce

$$p_s^{(k)} = \mathrm{Softmax}(W_{\mathrm{prob}} h_s + b_{\mathrm{prob}})_k, \quad c_s^{(k)} = \Psi_{\tau_s}(h_s + e^{(k)}), \quad N_s^{(k)} = \mathcal{I}_{\tau_s}(b_s, c_s^{(k)}). \tag{3}$$

Thus, $p_s^{(k)}$ and $N_s^{(k)}$ are sibling outputs of the same source representation: the former provides predictive prior mass, while the latter is the corresponding explicit physical hypothesis. Static map elements use a single mode.

The integrator $\mathcal{I}_{\tau_s}$ depends on the source type. Vehicle-like actors predict bounded acceleration and curvature, $a_t = a_{\max} \tanh \widetilde{a}_t$ and $\kappa_t = \kappa_{\max} \tanh \widetilde{\kappa}_t$, where $\kappa_t = \tan \delta_t / L$ follows the kinematic bicycle relation (Rajamani, 2006). Using midpoint discretization,

$$v_{t+1} = v_t + a_t \Delta t, \quad \theta_{t+1} = \theta_t + \bar{v}_t \kappa_t \Delta t, \quad \mathbf{p}_{t+1} = \mathbf{p}_t + \bar{v}_t \begin{bmatrix} \cos \bar{\theta}_t \\ \sin \bar{\theta}_t \end{bmatrix} \Delta t. \tag{4}$$

Holonomic actors instead use bounded planar-velocity integration, avoiding a vehicle-specific nonholonomic constraint, while static map elements are generated recursively from positive segment lengths and local heading increments. Complete state definitions and update equations for all three branches are provided in Appendices B.2–B.4.

Prior-weighted interaction. Each hypothesis is independently mapped to the common interaction space as $z_s^{(k)} = \mathcal{E}_{\tau_s}(N_s^{(k)})$. We set $K_{s,k}^h = W_K^h z_s^{(k)}$ and $V_{s,k}^h = W_V^h z_s^{(k)}$. For attention head h,

$$e_{t,s,k}^h = \frac{(Q_t^h)^\top K_{s,k}^h}{\sqrt{d_h}} + \log p_s^{(k)}, \quad \alpha_{t,s,k}^h = \mathrm{Softmax}_{s,k}(e_{t,s,k}^h), \quad m_t^h = \sum_{s,k} \alpha_{t,s,k}^h V_{s,k}^h. \tag{5}$$

This retains standard scaled dot-product attention (Vaswani et al., 2017). Predicted mode probabilities provide prior mass, while target-to-mediator compatibility performs target-dependent reweighting over valid source-mode pairs. The physical hypotheses therefore remain distinct until interaction rather than being averaged before the nonlinear attention operation.

Source-side no-bypass constraint. Proposition 1 additionally requires source-dependent information to enter the replaced interaction only through the mediator. Accordingly, the raw source query, raw source-anchor embedding, and additional source confidence are not reintroduced as keys, values, or continuous attention biases after mediator generation. If the host interaction uses source-side relative geometry, that quantity is recomputed from the corresponding $N_s^{(k)}$. Implementation details are in Appendices B.5–B.7; formal analyses are in Appendices A.4–A.6.

3.4 INSTANTIATION ACROSS INTERACTION PARADIGMS

CIB constrains the representation transmitted through selected interaction edges without changing the global computation topology of the host model. Across all configurations, the perception backbone, temporal self-interaction, target residual state, host task heads, active-source selection, and downstream decoding remain unchanged.

In the *sequential* configuration, source-side keys, values, positional terms, and mode priors in agent-to-ego and map-to-ego interactions are derived from the corresponding mediator representations. The planner therefore cannot directly consume raw source representations through the replaced edges. In the *parallel* configuration, ego-to-agent and agent-to-ego communication are implemented as separate directed CIB channels. Each direction constructs a source mediator before updating its

target. This replaces symmetric unconstrained latent exchange with explicit source-to-target mediation without claiming recovery of a real-world causal direction. In the *iterative* configuration, the mediator is regenerated at every refinement stage. Its physical state is used by the host anchor-update rule to construct the source anchor for the next stage, while the residual latent state is retained. Newly transmitted cross-entity information is therefore physically mediated at every stage without resetting the complete latent history. We refer to this mechanism as *stagewise physical re-anchoring*. Temporal alignment and cross-stage supervision are detailed in Appendices C.4–C.5.

3.5 Training and Physical Supervision

Structural mediation is useful only if the hypotheses transmitted through the bottleneck retain explicit physical semantics. We therefore preserve the host objectives and optimize $\mathcal{L} = \mathcal{L}_{\text{host}} + \lambda_{\text{phys}}\mathcal{L}_{\text{phys}}$, where the additional term directly supervises the integrated mediator states.

Let Ω^+ denote the CIB channel-stage pairs with valid supervision in the current minibatch. We define the aggregated physical-supervision objective over these valid pairs as

$$\mathcal{L}_{\text{phys}} = \frac{1}{|\Omega^+|} \sum_{(c,r)\in\Omega^+} \mathcal{L}^{c,r}_{\text{med}}, \qquad \mathcal{L}^{c,r}_{\text{med}} = \begin{cases} \mathcal{L}_{\text{state}} + \lambda_{\text{mode}}\mathcal{L}_{\text{mode}}, & \text{dynamic source,} \\ \mathcal{L}_{\text{map}}, & \text{static map source.} \end{cases} \tag{6}$$

Here $\mathcal{L}_{\text{state}}$ and $\mathcal{L}_{\text{map}}$ are normalized regression losses over valid integrated dynamic-state and map-state channels.

For a matched dynamic entity, the average positional displacement of mode k is

$$d_i^{(k)} = \frac{\sum_\tau m_{i,\tau} \left\| \widehat{\mathbf{p}}_{i,\tau}^{(k)} - \mathbf{p}_{i,\tau}^{\text{gt}} \right\|_2}{\sum_\tau m_{i,\tau}}. \tag{7}$$

Following the multiple-choice formulation of multimodal trajectory prediction (Chai et al., 2020), we select $k_i^\star = \arg\min_k d_i^{(k)}$ and use $\mathcal{L}_{\text{mode}} = -\log p_i^{(k_i^\star)}$ for its predicted probability.

Acceleration, curvature, steering, and planar velocity require no additional control annotations; they receive gradients through the differentiable physical rollout. The auxiliary objective is therefore not a causal-effect loss. Its role is to keep the representations transmitted through the interaction bottleneck tied to explicit kinematic or geometric states.

Dynamic and map supervision, temporal alignment, and cross-channel loss aggregation are detailed in Appendices C.2–C.5, with optimization details in Appendix C.6. Dynamic sources use six modes, static map elements use one mode, and $\Delta t = 0.5$ s.

4 Experiments

We first evaluate whether CIB provides consistent planning gains across interaction topologies and evaluation regimes. Matched architectural controls then separate generic bottleneck regularization, physical parameterization, and mediator supervision. Representation probes and paired simulator interventions examine whether the mediated representation exhibits the intended selective sensitivity to appearance and physical interaction changes. Finally, we compare with a related latent intervention method on matched backbones and transfer CIB to the stronger SparseDriveV2 planner.

4.1 Experimental Setup

We evaluate on nuScenes (Caesar et al., 2020), NAVSIM (Dauner et al., 2024), and Bench2Drive (Jia et al., 2024), covering three complementary regimes. On nuScenes, we report prefix-averaged trajectory displacement error, denoted as L2, and collision rate, denoted as CR. NAVSIM v1.1 is evaluated with the Predictive Driver Model Score, or PDMS, which combines collision avoidance, drivable-area compliance, time-to-collision, ego progress, and comfort. Reactive closed-loop evaluation uses Bench2Drive v0.0.4 with CARLA 0.9.15 and the official 220-route validation set; we report Driving Score, denoted as DS, and route Success Rate, denoted as SR. Benchmark versions and the three benchmark-specific metric definitions are provided in Appendices D.1–D.4, with statistical reporting in Appendix D.8.

Table 1: **Planning performance and controlled cross-paradigm evaluation.** Rows marked with † provide published nuScenes context under the standard 3-second planning protocol and thus cannot be used for strict paired claims. Unmarked SparseDrive–PhysDrive pairs are evaluated under our common stack; NAVSIM and Bench2Drive comparisons are therefore made only within these controlled pairs.

Method	Para.	nuScenes								NAVSIM	Bench2Drive	
		L2 (m) ↓				CR (%) ↓				PDMS ↑	DS ↑	SR (%) ↑
		1s	2s	3s	Avg.	1s	2s	3s	Avg.			
UniAD† (Hu et al., 2023)	S	0.45	0.70	1.04	0.73	0.62	0.58	0.63	0.61	83.4	45.81	16.36
VAD (Jiang et al., 2023)	S	0.34	0.59	0.92	0.62	0.21	0.34	0.58	0.38	81.2	44.35	16.91
SparseDrive-S	S	0.30	0.58	0.95	0.61	0.03	0.09	0.24	0.12	81.5	43.83	17.27
PhysDrive-S	S	**0.27**	**0.54**	**0.87**	**0.56**	**0.01**	**0.05**	**0.18**	**0.08**	**85.2**	**46.79**	**19.55**
MomAD† (Song et al., 2025)	P	0.31	0.57	0.91	0.60	0.01	0.05	0.22	0.09	-	45.35	17.44
BridgeAD† (Zhang et al., 2025)	P	0.28	0.55	0.92	0.58	**0.00**	**0.04**	0.20	0.08	-	-	-
SparseDrive-P (Sun et al., 2025)	P	0.28	0.55	0.91	0.58	**0.00**	0.06	0.18	0.08	82.8	44.93	18.64
PhysDrive-P	P	**0.26**	**0.52**	**0.85**	**0.54**	**0.00**	**0.04**	**0.15**	**0.06**	**86.4**	**48.36**	**21.36**
PPAD† (Chen et al., 2024)	I	0.31	0.56	0.87	0.58	0.08	0.12	0.38	0.19	-	-	-
VeteranAD† (Zhang et al., 2026)	I	0.33	0.59	0.94	0.62	0.03	0.12	0.34	0.16	**90.2**	**64.22**	**33.85**
SparseDrive-I	I	0.31	0.59	0.95	0.62	0.02	0.08	0.21	0.10	83.5	47.37	20.45
PhysDrive-I	I	**0.27**	**0.55**	**0.86**	**0.56**	**0.00**	**0.05**	**0.16**	**0.07**	88.0	50.72	22.27

† Published result. Literature rows provide context only because evaluation implementations are not identical to our frozen controlled stack.

For controlled comparisons, we instantiate sequential, parallel, and iterative interaction variants on the same SparseDrive host (Sun et al., 2025). Their CIB-equipped counterparts are denoted as PhysDrive-S, PhysDrive-P, and PhysDrive-I. Each paired model shares the same perception backbone, task heads, training horizon, source selection, and evaluation stack; the intervention is restricted to the specified cross-entity interaction channels and the associated mediator supervision.

The cross-paradigm study uses one complete training run per configuration to test consistency across architectures. The central matched-control study uses three independent training seeds and reports mean and standard deviation. Results quoted from prior work are marked explicitly and are not treated as strict paired comparisons unless they are rerun under our frozen evaluator.

4.2 Performance across Interaction Paradigms

Table 1 positions the performance results at two levels. Rows marked with † provide published nuScenes context from representative sequential, parallel, and iterative planners. Because these results originate from different implementation settings, we do not use them for strict cross-paper superiority claims. Our primary evidence instead comes from the three matched SparseDrive–PhysDrive comparisons evaluated under a common training and evaluation stack.

Across all three interaction topologies, CIB reduces nuScenes planning error and collision rate while improving NAVSIM PDMS. In the iterative configuration, average L2 decreases from 0.62 m to 0.56 m, while PDMS increases from 83.5 to 88.0. This pattern is consistent with the motivation that recurrent latent refinement can accumulate representation error, although the benchmark result alone does not establish that mechanism.

The improvement also extends to reactive closed-loop evaluation. Driving Score increases by approximately 3.0, 3.4, and 3.4 points for the sequential, parallel, and iterative configurations, respectively, while Success Rate improves by approximately 2.3, 2.7, and 1.8 percentage points. The consistent direction across open-loop, non-reactive, and closed-loop protocols indicates that the observed gain is not confined to a single interaction topology or evaluation regime. Taken together, these results support the cross-paradigm consistency of the performance gains, but should not by themselves be interpreted as attributing those gains to generic bottlenecking, explicit physical integration, or mediator supervision.

4.3 Matched Controls and Physical Parameterization

We first examine whether the improvement can be explained by a generic intermediate bottleneck. DSB retains the multimodal state representation, mediator supervision, physical-state re-encoder, and prior-weighted interaction but predicts mediator states directly. CIB without $\mathcal{L}_{\text{phys}}$ instead re-

Table 2: **Matched controls and physical parameterization.** Panel (a) reports mean±std over three independent training seeds on the parallel host, where Host denotes SparseDrive-P and Full CIB denotes PhysDrive-P. Panel (b) is a single-run diagnostic ablation and is not used to establish a statistical ranking among physical parameterizations.

(a) Matched controls

Method	L2 (m) ↓	CR (%) ↓	PDMS ↑
Host	0.580 ± 0.004	0.080 ± 0.006	82.8 ± 0.4
DSB	0.569 ± 0.005	0.077 ± 0.006	84.0 ± 0.5
CIB w/o phys. sup.	0.559 ± 0.005	0.071 ± 0.005	85.0 ± 0.5
Full CIB	0.546 ± 0.004	0.064 ± 0.005	86.4 ± 0.4

(b) Dynamic mediator

Parameterization	L2 (m) ↓	CR (%) ↓	PDMS ↑
Direct state	0.569	0.077	84.0
Accel. + yaw rate	0.554	0.070	85.8
Accel. + curvature	0.550	0.067	86.1
Class-aware	**0.546**	**0.064**	**86.4**

tains the physical integrator while removing explicit mediator supervision, thereby separating physical integration from mediator supervision under the same interaction structure.

Panel (a) of Table 2 shows that DSB already improves over the host, confirming that the constrained intermediate representation itself provides a useful regularization effect. Explicit physical integration yields a further improvement, and the complete CIB achieves the strongest mean performance. The mean gaps across the three-seed study are also larger than the observed run-to-run dispersion, although we do not interpret the reported standard deviations as a formal significance test.

Panel (b) is a diagnostic parameterization ablation under a fixed training setting. Direct state corresponds to the DSB representation, followed by yaw-rate, curvature, and class-aware physical parameterizations. Explicit integration improves over direct-state prediction, while the curvature and class-aware variants provide the strongest overall single-run trade-off. Since the differences among the physical variants are small and this panel does not report multiple training seeds, we use it to establish that the choice of physical representation matters rather than to claim a statistically resolved ranking among parameterizations.

4.4 Controlled Mechanism Validation

Matched controls establish that the physical representation contributes to performance, but they do not reveal how information accessibility changes inside the interaction mechanism. We therefore examine the mechanism at both the representation and behavioral levels.

Representation probing. After freezing the driving model, we train low-capacity linear probes on several internal representations. For multimodal mediators, we probe the probability-weighted physical and encoded summaries $\overline{N}_s = \sum_k p_s^{(k)} N_s^{(k)}$ and $\overline{z}_s = \sum_k p_s^{(k)} z_s^{(k)}$; these summaries are diagnostic only and are not used by the forward interaction. Appearance decodability is measured by balanced accuracy, denoted as BA, over four equally represented rendering conditions; chance performance is therefore 25%. To ensure that a reduction in appearance decodability is not caused by generic information loss, we simultaneously recover relative position, speed, and heading and report their normalized mean absolute error, denoted as nMAE. Lower appearance BA is interpreted as desirable only when physical nMAE remains competitive.

Table 3 shows that the raw query retains substantial appearance information even after a dimension-matched random projection. DSB reduces this decodability, while the CIB physical summary reduces it further and simultaneously improves physical-state recovery. The encoded mediator exhibits the same overall pattern. These results characterize linear decodability and are not interpreted as estimates of mutual information.

Paired interventions. We next evaluate planning behavior under controlled scene pairs. Nuisance Intervention Sensitivity, or NIS, is the average trajectory change when only the rendering condition varies and the physical scene is held fixed. Reaction Sensitivity, or RS, measures the trajectory response when a behaviorally relevant agent is physically changed. False Braking Rate, or FBR, measures the fraction of eligible agent-removal cases in which the planner still produces strong unnecessary braking. Unlike NIS and FBR, RS has no universal monotonic direction and is interpreted only together with the intervention semantics.

PhysDrive-P substantially reduces NIS relative to SparseDrive-P while producing a larger response to removal of a relevant actor and a lower FBR. The joint pattern indicates that reduced sensitivity to

Table 3: **Controlled mechanism validation.** Appearance BA measures linear decodability and is interpreted jointly with physical nMAE. NIS measures sensitivity to appearance-only changes; removal RS measures response magnitude to removing a relevant actor and has no preferred direction; FBR measures false braking on eligible removal cases.

(a) Representation probe

Representation	Appearance BA (%)	Physical nMAE ↓
Raw query	88.6	0.112
Random projection	79.4	0.126
DSB state	66.8	0.095
Physical summary	**55.8**	**0.081**
Encoded summary	59.6	0.084

(b) Paired interventions

Method	NIS (m) ↓	Removal RS (m)	FBR (%) ↓
SparseDrive-P	0.21	0.16	27.1
+DSB	0.16	0.35	15.6
PhysDrive-P	**0.09**	**0.61**	**7.2**

controlled appearance changes is not explained by uniformly suppressing environmental response. We refer to this behavior as *selective sensitivity*. Full probe and intervention protocols are provided in Appendices E.2–E.4, with metric definitions in Appendix D.5. Additional per-condition and speed-intervention results are reported in Appendix F.3.

4.5 External Comparisons and Architectural Transfer

CausalVAD provides a strong same-backbone reference through its Sparse Causal Intervention Scheme, or SCIS (Tang et al., 2026). On the parallel and iterative SparseDrive hosts, SCIS reports $0.53/0.06$ and $0.55/0.07$ L2/CR, respectively. CIB operates in a comparable range but is not uniformly stronger on open-loop metrics. The two methods intervene at different points in the representation pipeline: SCIS adjusts latent context using learned prototypes, whereas CIB constrains the source representation transmitted through cross-entity interaction. Their composition provides a modest additional gain under the tested hosts, which we interpret as empirical complementarity rather than a ranking of causal validity.

We further insert CIB into the scene-to-planning interaction interfaces of SparseDriveV2 (Sun et al., 2026), leaving its trajectory vocabulary, candidate scoring, and trajectory-to-control pipeline unchanged. SparseDriveV2 reports 89.15 DS and 70.00% SR on Bench2Drive; our reproduction obtains $89.0/69.5$, and adding CIB yields $89.9/71.5$. Since SparseDriveV2 is already substantially stronger than the host architectures used in our primary controlled study, this result primarily supports architectural portability rather than a stronger front-door interpretation. Same-backbone composition, controlled-intervention comparison, and SparseDriveV2 implementation details are provided in Appendices G.2– G.4.

5 Conclusion

We study architectural shortcut pathways in end-to-end autonomous driving and introduce CIB to constrain source-side information flow in selected cross-entity interactions through explicit kinematic and geometric mediators. At the network level, this design provides source-side structural mediation and conditional information non-increase under deterministic mediation, while prior-weighted attention and discrete physical integration admit query-conditioned and bounded interpretations. Across three interaction paradigms, matched controls indicate that the performance gains are not fully explained by generic bottleneck regularization. Representation probes and paired interventions further show lower sensitivity to appearance-only variation while preserving responses to behaviorally relevant physical changes. Together, these results support physically mediated communication as a useful structural inductive bias for robust and analyzable end-to-end driving. The approach nevertheless has clear boundaries. Restricting interaction content to the current physical mediators may attenuate semantic cues not fully captured by kinematic or geometric states, such as turn signals or pedestrian gestures. The discrete kinematic model also omits full vehicle-dynamics effects such as tire slip, friction limits, and actuator dynamics. Moreover, structural mediation does not imply independence from all latent environmental factors or establish classical front-door causal identification. Retaining multiple physical hypotheses until attention also increases memory and computation. Richer mediator semantics, higher-fidelity dynamics, and more efficient multimodal interaction remain promising directions for future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with refining the conceptual framing, discussing and checking mathematical claims and proof drafts, critically reviewing experimental designs and ablation plans, and supporting literature search, hyperparameter discussion, manuscript organization, and language editing. All AI-assisted material was reviewed by the authors. Mathematical claims and proofs were independently checked by the authors, methodological and experimental choices were verified against primary literature, implementation details, and dataset specifications, and citations were checked against their original sources. Generative AI tools did not autonomously determine the final scientific claims or select which experimental results to report. The authors take full responsibility for the entire content of this work.

REPRODUCIBILITY STATEMENT

We provide reproducibility information at the theoretical, implementation, and evaluation levels. Section 3 specifies the complete CIB computation and its core assumptions, while Appendix A provides the formal statements, assumptions, and proofs. Appendices B and C detail the physical integrators, network architecture, initialization, training objectives, temporal alignment, and optimization hyperparameters. Appendix D specifies the benchmark versions, evaluation protocols, and diagnostic metrics, while Appendix E documents the representation probes and controlled simulator interventions. Extended ablations, statistical analyses, and cross-method and cross-architecture experiments are reported in Appendices F and G. Controlled comparisons use fixed evaluation protocols and matched optimization budgets, with independent training seeds for the central matched-control study.

REFERENCES

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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pp. 11621–11631, 2020.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, pp. 213–229. Springer, 2020.

Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In *Proceedings of the Conference on Robot Learning*, pp. 86–99. PMLR, 2020.

Zhili Chen, Maosheng Ye, Shuangjie Xu, Tongyi Cao, and Qifeng Chen. Ppad: Iterative interactions of prediction and planning for end-to-end autonomous driving. In *Eur. Conf. Comput. Vis.*, pp. 239–256. Springer, 2024.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Adv. Neural Inform. Process. Syst.*, 37:28706–28719, 2024.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pp. 1–16, 2017.

Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24823–24834, 2025.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Mach. Intell.*, 2(11):665–673, 2020.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pp. 17853–17862, 2023.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Adv. Neural Inform. Process. Syst.*, 37:819–844, 2024.

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 *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 8340–8350, 2023.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Zhenxin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Zuxuan Wu, and Jose M. Alvarez. Hydra-next: Robust closed-loop driving with open-loop training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27305–27314, 2025.

Haicheng Liao, Chengyue Wang, Zhenning Li, Yongkang Li, Bonan Wang, Guofa Li, and Chengzhong Xu. Physics-informed trajectory prediction for autonomous driving under missing observation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 6841–6849, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Ignavier Ng, Shaoan Xie, Xinshuai Dong, Peter Spirtes, and Kun Zhang. Causal representation learning from general environments under nonparametric mixing. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, pp. 3700–3708, 2025.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Rajesh Rajamani. *Vehicle dynamics and control*. Springer, 2006.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proc. IEEE*, 109 (5):612–634, 2021.

Ziying Song, Caiyan Jia, Lin Liu, Hongyu Pan, Yongchang Zhang, Junming Wang, Xingyu Zhang, Shaoqing Xu, Lei Yang, and Yadan Luo. Don't shake the wheel: Momentum-aware planning in end-to-end autonomous driving. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pp. 22432–22441, 2025.

Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. In *2025 IEEE Int. Conf. Robot. Autom.*, pp. 8795–8801. IEEE, 2025.

Wenchao Sun, Xuewu Lin, Keyu Chen, Zixiang Pei, Xiang Li, Yining Shi, and Sifa Zheng. Sparsedrivev2: Scoring is all you need for end-to-end autonomous driving. In *European Conference on Computer Vision*, pp. 446–463. Springer, 2026.

Jiacheng Tang, Zhiyuan Zhou, Zhuolin He, Jia Zhang, Kai Zhang, and Jian Pu. Causalvad: Deconfounding end-to-end autonomous driving via causal intervention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32124–32133, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

Wenyuan Zeng, Wenjie Luo, Simon Suo, Abbas Sadat, Bin Yang, Sergio Casas, and Raquel Urtasun. End-to-end interpretable neural motion planner. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pp. 8660–8669, 2019.

Bozhou Zhang, Nan Song, Xin Jin, and Li Zhang. Bridging past and future: End-to-end autonomous driving with historical prediction and planning. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pp. 6854–6863, 2025.

Bozhou Zhang, Jingyu Li, Nan Song, and Li Zhang. Perception in plan: Coupled perception and planning for end-to-end autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12376–12384, 2026.

A Formal Analysis of Physical Mediation

This section provides a formal analysis of the Causal Interaction Bottleneck introduced in Section 3. Our objective is not to establish a front-door identification theorem for causal effects in the physical world. Instead, we characterize the structural constraints imposed by CIB on the network computation graph and derive the properties that follow from these constraints. We first clarify the relation between CIB and the classical front-door criterion and then establish source-side structural mediation and its information-theoretic consequence. We next show that prior-weighted mode-expanded attention admits an exact expectation representation under a query-conditioned interaction distribution and characterize the stability of this normalized attention functional under Wasserstein perturbations of the predictive mediator distributions. Finally, we derive explicit bounds for the curvature-form discrete kinematic update and discuss the scope of stagewise mediation in iterative refinement.

A.1 Relation to Classical Front-Door Identification

Consider a classical causal system with treatment X, mediator M, and outcome Y. The front-door criterion requires that M intercept all directed paths from X to Y, that no unblocked back-door path exists from X to M, and that all back-door paths from M to Y are blocked by conditioning on X (Pearl, 2009). Under these conditions and the corresponding positivity assumptions, the causal effect for discrete variables is identified as

$$P(Y = y \mid do(X = x)) = \sum_{m} P(m \mid x) \sum_{x'} P(y \mid m, x')P(x'). \tag{8}$$

CIB does not assume that learned source representations, physical mediators, and target representations satisfy these identification conditions. In particular, the physical mediator is generated from a learned source representation and can therefore remain statistically dependent on environmental context. Compressing a latent representation into a constrained physical state is not sufficient to establish independence from an unobserved confounder. We consequently use the front-door criterion only as structural motivation for mediated information flow rather than interpreting CIB as a neural implementation of Equation 8.

Let

$$\chi_s = (q_s, b_s, \tau_s) \tag{9}$$

denote the complete descriptor of source s, where q_s is its latent query, b_s is the physical anchor required by the corresponding integrator, and τ_s is its entity type. CIB constructs

$$\mathcal{M}_s = \mathcal{G}(\chi_s) = \left(\tau_s, \left\{N_s^{(k)}, p_s^{(k)}\right\}_{k=1}^{K_s}\right). \tag{10}$$

For target t, define

$$\mathbf{M}_t = \{\mathcal{M}_s : s \in \mathcal{S}_t\} \tag{11}$$

as the collection of mediator representations from all valid sources. The CIB update can then be written abstractly as

$$m_t = \mathcal{H}\left(q_t, \mathbf{M}_t\right), \quad q_t^+ = \mathcal{F}\left(q_t, m_t\right). \tag{12}$$

The defining implementation constraint is that, once the active source set and attention mask are fixed, no source-specific quantity can enter $\mathcal{H}$ except through its corresponding mediator representation $\mathcal{M}_s$.

A.2 Structural Mediation and the Source-Side No-Bypass Property

Proof of Proposition 1. Fix the target query q_t, the active source set and attention mask, and all mediator representations except that of source s. Consider two source descriptors χ_s and $\widetilde{\chi}_s$ satisfying

$$\mathcal{G}(\chi_s) = \mathcal{G}(\widetilde{\chi}_s). \tag{13}$$

The equality includes the entity type, all physical hypotheses, and their mode probabilities. The type-specific encoder therefore produces identical source-mode embeddings, and the corresponding keys, values, and mode priors are unchanged. Since every other source mediator is fixed, the complete set of attention logits in Equation 5 is identical. The resulting attention distribution and multi-head message are consequently identical. Applying the same target update gives

$$q_t^+(\chi_s) = q_t^+(\widetilde{\chi}_s). \tag{14}$$

Hence, source-side variation that leaves the complete mediator representation unchanged cannot affect the target through the CIB interaction. □

The result follows entirely from the computation graph and does not require a causal-identification assumption. It applies specifically to cross-entity interaction edges replaced by CIB. Information already stored in the target query is not removed by the proposition.

A.3 Information Non-Increase under Physical Mediation

Let Z denote a contextual variable and define

$$B_s = (b_s, \tau_s), \quad \chi_s = (q_s, B_s). \tag{15}$$

Because the complete mediator representation is deterministically generated from χ_s,

$$\mathcal{M}_s = \mathcal{G}(\chi_s), \tag{16}$$

the variables satisfy the Markov relation

$$Z \longrightarrow \chi_s \longrightarrow \mathcal{M}_s. \tag{17}$$

Proposition 2 (Information non-increase)**.** *Whenever the relevant mutual informations are well defined,*

$$I(Z; \mathcal{M}_s) \le I(Z; \chi_s). \tag{18}$$

Moreover, conditional on the source physical anchor and entity type,

$$I(Z; \mathcal{M}_s \mid B_s) \le I(Z; q_s \mid B_s). \tag{19}$$

Proof. Equation 18 follows from the data processing inequality for the Markov chain $Z \to \chi_s \to \mathcal{M}_s$. For fixed B_s, the mediator is a deterministic function of (q_s, B_s) and therefore satisfies

$$\mathcal{M}_s \perp Z \mid (q_s, B_s). \tag{20}$$

Applying the data processing inequality to the conditional distribution given B_s gives Equation 19 (Cover & Thomas, 2006). □

Proposition 2 is an information upper bound, not an isolation result. It neither implies $I(Z; \mathcal{M}_s \mid B_s) = 0$ nor guarantees a strict reduction. In particular, the mode probabilities are part of $\mathcal{M}_s$ and may retain contextual information from the source query. The amount of nuisance information that remains accessible is therefore evaluated empirically.

A.4 Prior-Weighted Attention as a Query-Conditioned Expectation

Fix a target query q_t and attention head h. Let

$$\Xi_t = \{(s,k) : s \in \mathcal{S}_t,\ 1 \le k \le K_s\} \tag{21}$$

denote the finite index set of valid source-mode pairs. The predicted mode probabilities induce the base distribution

$$\rho_t(s,k) = \frac{p_s^{(k)}}{|\mathcal{S}_t|}. \tag{22}$$

Since $\sum_{k=1}^{K_s} p_s^{(k)} = 1$ for every source, ρ_t is a probability distribution over Ξ_t.

For head h, define the target-to-token compatibility

$$\ell_{t,h}(s,k) = \frac{(Q_t^h)^\top K_{s,k}^h}{\sqrt{d_h}}. \tag{23}$$

The target query induces the reweighted distribution

$$\Pi_{t,h}(s,k) = \frac{\rho_t(s,k)\exp\left(\ell_{t,h}(s,k)\right)}{\sum_{s',j}\rho_t(s',j)\exp\left(\ell_{t,h}(s',j)\right)}. \tag{24}$$

Proposition 3 (Exact query-conditioned attention representation)**.** *The prior-weighted mode-expanded attention in Section 3.3 satisfies*

$$\Pi_{t,h}(s,k) = \alpha_{t,s,k}^h, \tag{25}$$

and the output of head h can be written exactly as

$$m_t^h = \mathbb{E}_{(S,K)\sim\Pi_{t,h}}\left[V_{S,K}^h\right]. \tag{26}$$

Proof. Substituting Equation 22 into Equation 24 gives

$$\Pi_{t,h}(s,k) = \frac{p_s^{(k)}\exp\left(\ell_{t,h}(s,k)\right)/|\mathcal{S}_t|}{\sum_{s',j}p_{s'}^{(j)}\exp\left(\ell_{t,h}(s',j)\right)/|\mathcal{S}_t|}. \tag{27}$$

The common factor $|\mathcal{S}_t|^{-1}$ cancels, yielding

$$\Pi_{t,h}(s,k) = \frac{p_s^{(k)}\exp\left(\ell_{t,h}(s,k)\right)}{\sum_{s',j}p_{s'}^{(j)}\exp\left(\ell_{t,h}(s',j)\right)} = \alpha_{t,s,k}^h. \tag{28}$$

Equation 26 then follows directly from the definition of the attention output as the weighted sum of its value tokens. □

The predicted mode probabilities therefore define prior mass over physical hypotheses, while target-to-mediator compatibility performs a query-dependent exponential reweighting. Proposition 3 is an exact representation of the network computation. We do not interpret $\Pi_{t,h}$ as a causal posterior, nor do we identify this expectation with the marginalization in the classical front-door functional.

A.5 Stability under Mediator-Distribution Approximation

We next analyze the sensitivity of the attention message to approximation of the source-specific physical mediator distributions. The result concerns the stability of the network computation and does not assume that any reference distribution is causally identified.

Fix a target t and attention head h. For each source $s \in \mathcal{S}_t$, let $(\mathcal{N}_{\tau_s}, d_s)$ denote the physical mediator space associated with its entity type, and let

$$\mu_s, \nu_s \in \mathcal{P}_1\left(\mathcal{N}_{\tau_s}\right) \tag{29}$$

be two probability measures with finite first moments. We define the source-averaged Wasserstein discrepancy (Villani, 2009) as

$$\overline{W}_1(\boldsymbol{\mu}, \boldsymbol{\nu}) = \frac{1}{|\mathcal{S}_t|} \sum_{s \in \mathcal{S}_t} W_1(\mu_s, \nu_s), \tag{30}$$

where $\boldsymbol{\mu} = \{\mu_s\}_{s \in \mathcal{S}_t}$ and $\boldsymbol{\nu} = \{\nu_s\}_{s \in \mathcal{S}_t}$.

The finite distribution predicted by CIB for source s is

$$\widehat{\mu}_s = \sum_{k=1}^{K_s} p_s^{(k)} \delta_{N_s^{(k)}}. \tag{31}$$

For a physical state $n \in \mathcal{N}_{\tau_s}$, define the value function

$$v_{s,h}(n) = W_V^h \mathcal{E}_{\tau_s}(n), \tag{32}$$

the target-to-mediator compatibility

$$\ell_{t,s,h}(n) = \frac{(Q_t^h)^\top W_K^h \mathcal{E}_{\tau_s}(n)}{\sqrt{d_h}}, \tag{33}$$

and the positive compatibility weight

$$w_{t,s,h}(n) = \exp(\ell_{t,s,h}(n)). \tag{34}$$

The attention functional induced by the collection of source distributions $\boldsymbol{\mu}$ is

$$\mathcal{T}_{t,h}(\boldsymbol{\mu}) = \frac{\frac{1}{|\mathcal{S}_t|} \sum_{s \in \mathcal{S}_t} \int w_{t,s,h}(n) v_{s,h}(n) \, d\mu_s(n)}{\frac{1}{|\mathcal{S}_t|} \sum_{s \in \mathcal{S}_t} \int w_{t,s,h}(n) \, d\mu_s(n)}. \tag{35}$$

When $\mu_s = \widehat{\mu}_s$ for every source, Equation 35 reduces exactly to the CIB attention output m_t^h.

Assume that, on the supports of the distributions under consideration, there exist constants $w_{\min} > 0$, $w_{\max}$, M_v, L_w, and L_v that hold uniformly over all $s \in \mathcal{S}_t$:

$$w_{\min} \le w_{t,s,h}(n) \le w_{\max}, \quad \|v_{s,h}(n)\|_2 \le M_v, \tag{36}$$

and

$$|w_{t,s,h}(n) - w_{t,s,h}(n')| \le L_w d_s(n, n'), \tag{37}$$

$$\|v_{s,h}(n) - v_{s,h}(n')\|_2 \le L_v d_s(n, n'). \tag{38}$$

Proposition 4 (Stability of prior-weighted attention)**.** *Under the assumptions above,*

$$\|\mathcal{T}_{t,h}(\boldsymbol{\mu}) - \mathcal{T}_{t,h}(\boldsymbol{\nu})\|_2 \le C_{\text{att}} \, \overline{W}_1(\boldsymbol{\mu}, \boldsymbol{\nu}), \tag{39}$$

where

$$C_{\text{att}} = \frac{w_{\max} L_v + M_v L_w}{w_{\min}} + \frac{w_{\max} M_v L_w}{w_{\min}^2}. \tag{40}$$

Proof. Define

$$A_{\boldsymbol{\mu}} = \frac{1}{|\mathcal{S}_t|} \sum_s \int w_{t,s,h}(n) v_{s,h}(n) \, d\mu_s(n) \tag{41}$$

and

$$Z_{\boldsymbol{\mu}} = \frac{1}{|\mathcal{S}_t|} \sum_s \int w_{t,s,h}(n) \, d\mu_s(n), \tag{42}$$

with $A_{\boldsymbol{\nu}}$ and $Z_{\boldsymbol{\nu}}$ defined analogously.

For any source s, the product $w_{t,s,h}(n)v_{s,h}(n)$ is Lipschitz because

$$\begin{aligned}&\|w(n)v(n) - w(n')v(n')\|_2\\&\leq w(n)\,\|v(n) - v(n')\|_2 + \|v(n')\|_2\,|w(n) - w(n')|\\&\leq (w_{\max}L_v + M_vL_w)\,d_s(n,n'),\end{aligned} \tag{43}$$

where the target, head, and source subscripts are omitted for readability. Applying the coupling definition of W_1 to each source and averaging gives

$$\|A_{\boldsymbol{\mu}} - A_{\boldsymbol{\nu}}\|_2 \leq (w_{\max}L_v + M_vL_w)\,\overline{W}_1\,(\boldsymbol{\mu},\boldsymbol{\nu})\,. \tag{44}$$

Similarly,

$$|Z_{\boldsymbol{\mu}} - Z_{\boldsymbol{\nu}}| \leq L_w\overline{W}_1\,(\boldsymbol{\mu},\boldsymbol{\nu})\,. \tag{45}$$

The lower bound on w yields

$$Z_{\boldsymbol{\mu}} \geq w_{\min}, \quad Z_{\boldsymbol{\nu}} \geq w_{\min}, \tag{46}$$

and the bounded-value assumption gives

$$\|A_{\boldsymbol{\nu}}\|_2 \leq w_{\max}M_v. \tag{47}$$

Therefore,

$$\begin{aligned}&\|\mathcal{T}_{t,h}\,(\boldsymbol{\mu}) - \mathcal{T}_{t,h}\,(\boldsymbol{\nu})\|_2\\&\leq \frac{\|A_{\boldsymbol{\mu}} - A_{\boldsymbol{\nu}}\|_2}{Z_{\boldsymbol{\mu}}} + \|A_{\boldsymbol{\nu}}\|_2\,\frac{|Z_{\boldsymbol{\mu}} - Z_{\boldsymbol{\nu}}|}{Z_{\boldsymbol{\mu}}Z_{\boldsymbol{\nu}}},\end{aligned} \tag{48}$$

which yields Equation 39. □

Proposition 4 is a conditional stability result. It does not assert that the compatibility or value functions are globally bounded and Lipschitz. The assumptions are required only on the supports of the mediator distributions being compared.

In particular, let $\mu_s^\star$ denote a reference predictive distribution for source s. If a sequence of finite CIB distributions $\widehat{\mu}_{s,K}$ satisfies

$$\frac{1}{|\mathcal{S}_t|}\sum_s W_1\,(\widehat{\mu}_{s,K}, \mu_s^\star) \longrightarrow 0, \tag{49}$$

then

$$\mathcal{T}_{t,h}\,(\widehat{\boldsymbol{\mu}}_K) \longrightarrow \mathcal{T}_{t,h}\,(\boldsymbol{\mu}^\star)\,. \tag{50}$$

The reference distributions are predictive mediator distributions and are not interpreted as identified causal distributions.

A.6 Bounds of Curvature-Form Kinematic Integration

For vehicle mediators, CIB predicts bounded acceleration and curvature,

$$a_t = a_{\max}\tanh\widetilde{a}_t, \quad \kappa_t = \kappa_{\max}\tanh\widetilde{\kappa}_t. \tag{51}$$

The curvature parameter follows the kinematic bicycle relation

$$\dot{\theta} = v\kappa, \quad \kappa = \frac{\tan\delta}{L}, \tag{52}$$

where δ is the steering angle and L is the wheelbase (Rajamani, 2006).

The discrete update used by CIB is

$$v_{t+1} = v_t + a_t\Delta t, \quad \bar{v}_t = \frac{v_t + v_{t+1}}{2}, \tag{53}$$

$$\theta_{t+1} = \theta_t + \bar{v}_t\kappa_t\Delta t, \tag{54}$$

and

$$\mathbf{p}_{t+1} = \mathbf{p}_t + \bar{v}_t\begin{bmatrix}\cos\bar{\theta}_t\\ \sin\bar{\theta}_t\end{bmatrix}\Delta t, \quad \bar{\theta}_t = \frac{\theta_t + \theta_{t+1}}{2}. \tag{55}$$

Proposition 5 (Bounded discrete kinematic update)**.** *Let $\Delta t > 0$. The updates in Equations 51–55 satisfy*

$$|a_t| \le a_{\max}, \quad |\kappa_t| \le \kappa_{\max}, \tag{56}$$

and

$$|v_{t+1} - v_t| \le a_{\max}\Delta t. \tag{57}$$

Define the discrete displacement length

$$\Delta\ell_t = \|\mathbf{p}_{t+1} - \mathbf{p}_t\|_2 . \tag{58}$$

Then

$$\Delta\ell_t = |\bar{v}_t|\Delta t \tag{59}$$

and

$$|\theta_{t+1} - \theta_t| \le \kappa_{\max}\Delta\ell_t. \tag{60}$$

Whenever $\Delta\ell_t > 0$,

$$\frac{|\theta_{t+1} - \theta_t|}{\Delta\ell_t} = |\kappa_t| \le \kappa_{\max}. \tag{61}$$

If $\Delta\ell_t = 0$, then $\theta_{t+1} = \theta_t$.

Proof. Since $|\tanh x| \le 1$, Equation 51 gives

$$|a_t| \le a_{\max}, \quad |\kappa_t| \le \kappa_{\max}. \tag{62}$$

Equation 53 therefore yields

$$|v_{t+1} - v_t| = |a_t|\Delta t \le a_{\max}\Delta t. \tag{63}$$

The direction vector in Equation 55 has unit Euclidean norm, so

$$\Delta\ell_t = |\bar{v}_t|\Delta t. \tag{64}$$

Using Equation 54,

$$|\theta_{t+1} - \theta_t| = |\bar{v}_t||\kappa_t|\Delta t \tag{65}$$

$$= |\kappa_t|\Delta\ell_t \tag{66}$$

$$\le \kappa_{\max}\Delta\ell_t. \tag{67}$$

For $\Delta\ell_t > 0$, division by $\Delta\ell_t$ gives the stated heading-change bound. If $\Delta\ell_t = 0$, then $\bar{v}_t = 0$, and Equation 54 gives $\theta_{t+1} = \theta_t$. □

Proposition 5 characterizes the bounds enforced by the adopted discrete kinematic update. It does not establish full vehicle dynamic feasibility. In particular, it does not model tire slip, friction limits, steering-rate limits, load transfer, or actuator dynamics. The update also does not impose a nonnegative-speed constraint unless such a constraint is enforced separately. We therefore interpret the result as a bound on the discrete kinematic representation rather than as a guarantee of complete physical feasibility.

A.7 Stagewise Mediation in Iterative Refinement

At refinement stage r, define the source descriptor as

$$\chi_s^{(r)} = \left(q_s^{(r)}, b_s^{(r)}, \tau_s\right), \tag{68}$$

with mediator representation

$$\mathcal{M}_s^{(r)} = \mathcal{G}_r\left(\chi_s^{(r)}\right). \tag{69}$$

For target t, the interaction message is

$$m_t^{(r)} = \mathcal{H}_r\left(q_t^{(r)}, \left\{\mathcal{M}_s^{(r)}\right\}_{s\in\mathcal{S}_t^{(r)}}\right), \tag{70}$$

followed by

$$q_t^{(r+1)} = \mathcal{F}_r\left(q_t^{(r)}, m_t^{(r)}\right). \tag{71}$$

Each refinement stage therefore satisfies the same source-side mediation property as Proposition 1: once the active source set and mask are fixed, newly transmitted source information enters the target only through the corresponding mediator representations.

The mediator-derived physical state is additionally used by the host-specific deterministic anchor-update rule to construct $b_s^{(r+1)}$ for the next refinement stage. This repeatedly re-anchors spatial reasoning to an explicit kinematic or geometric state. The residual latent query, however, is retained and can carry information from earlier stages. CIB therefore does not erase the complete latent history and does not implement a causal reset or a *do* intervention. We describe this mechanism as *stagewise physical re-anchoring*.

Overall, the guarantees in this appendix concern the computation graph and the adopted physical parameterization. They establish compulsory mediation of source-to-target communication, information non-increase under deterministic mediation, exact expectation under the query-conditioned interaction distribution, and bounded discrete kinematics. They do not establish independence from every latent environmental factor or identify a real-world interventional causal effect.

B Detailed CIB Architecture and Physical Integrators

This section specifies the network architecture and forward computation of the Causal Interaction Bottleneck. We describe how latent source queries are converted into physical mediators, how the resulting state sequences are re-encoded as interaction tokens, and how the source-side no-bypass constraint is enforced in implementation. CIB replaces cross-entity attention channels while preserving the perception backbone, the residual state of the target query, and the temporal identity mechanisms of the host architecture. All continuous operations remain differentiable and receive gradients from the downstream driving objectives.

B.1 Overall Forward Architecture

Consider source descriptors

$$\chi_s = (q_s, b_s, \tau_s), \quad s = 1, \dots, S, \tag{72}$$

where $q_s \in \mathbb{R}^d$ is the latent source query, b_s is the physical anchor used to initialize the corresponding rollout, and τ_s identifies the entity type. The complete generator $\mathcal{G}$ comprises a neural physical-parameter predictor followed by a type-specific deterministic integrator.

The latent query is first processed by a shared generator trunk,

$$h_s = \Phi_{\mathrm{G}}(q_s), \tag{73}$$

where Φ_{G} is a two-layer MLP with hidden dimension d and ReLU activations. For dynamic entities with multimodal futures, each mode is associated with a learned embedding

$$e^{(k)} \in \mathbb{R}^d, \quad h_s^{(k)} = h_s + e^{(k)}. \tag{74}$$

When the host architecture already provides mode-specific motion embeddings, we reuse those embeddings instead.

The mode-scoring head produces logits

$$\beta_s = W_{\mathrm{prob}} h_s + b_{\mathrm{prob}}, \tag{75}$$

with probabilities

$$p_s^{(k)} = \mathrm{Softmax}(\beta_s)_k. \tag{76}$$

For each mode, a type-specific physical head predicts a control or geometric parameter sequence

$$\mathbf{c}_s^{(k)} = \Psi_{\tau_s}\left(h_s^{(k)}\right). \tag{77}$$

The corresponding physical integrator is initialized by b_s and produces

$$N_s^{(k)} = \mathcal{I}_{\tau_s}\left(b_s, \mathbf{c}_s^{(k)}\right). \tag{78}$$

Vehicle-like actors use the curvature-based nonholonomic integrator, holonomic actors use bounded planar integration, and static map elements use the intrinsic polyline integrator. The complete mediator representation is therefore

$$\mathcal{M}_s = \left(\tau_s, \left\{ N_s^{(k)}, p_s^{(k)} \right\}_{k=1}^{K_s} \right). \tag{79}$$

We use $K_s = 6$ for dynamic entities, matching the multimodal motion representation of the host architecture, and $K_s = 1$ for static map elements. The temporal support likewise follows the host task: dynamic-agent mediators use 12 future steps at $\Delta t = 0.5$ s, ego mediators use 6 future steps, and map mediators use the same 20-point polyline representation as the host mapping head.

Finally, each physical hypothesis is independently mapped to the common interaction dimension by the encoder associated with its entity type,

$$z_s^{(k)} = \mathcal{E}_{\tau_s} \left(N_s^{(k)} \right). \tag{80}$$

All source-mode representations remain distinct until the prior-weighted attention stage.

B.2 Curvature-Based Integration for Vehicle-Like Agents

For a vehicle-like source, the physical anchor in the source descriptor is

$$b_s = (x_0, y_0, v_0, \theta_0). \tag{81}$$

For each physical mode k, the corresponding control head predicts

$$\left\{ \widetilde{a}_\tau^{(k)}, \widetilde{\kappa}_\tau^{(k)} \right\}_{\tau=0}^{T-1}. \tag{82}$$

The logits are mapped to bounded acceleration and curvature,

$$a_\tau^{(k)} = a_{\max} \tanh \widetilde{a}_\tau^{(k)}, \quad \kappa_\tau^{(k)} = \kappa_{\max} \tanh \widetilde{\kappa}_\tau^{(k)}. \tag{83}$$

The curvature follows the kinematic bicycle relation $\kappa = \tan \delta / L$, where δ is the steering angle and L is the wheelbase (Rajamani, 2006).

Starting from b_s, the longitudinal state is updated as

$$v_{\tau+1} = v_\tau + a_\tau \Delta t, \quad \bar{v}_\tau = \frac{v_\tau + v_{\tau+1}}{2}, \tag{84}$$

followed by

$$\theta_{\tau+1} = \theta_\tau + \bar{v}_\tau \kappa_\tau \Delta t. \tag{85}$$

Using

$$\bar{\theta}_\tau = \frac{\theta_\tau + \theta_{\tau+1}}{2}, \tag{86}$$

the spatial state is integrated as

$$\begin{aligned} x_{\tau+1} &= x_\tau + \bar{v}_\tau \cos \bar{\theta}_\tau \Delta t, \\ y_{\tau+1} &= y_\tau + \bar{v}_\tau \sin \bar{\theta}_\tau \Delta t. \end{aligned} \tag{87}$$

The state supplied to the physical encoder is

$$n_\tau^{(k)} = \left[x_\tau, y_\tau, v_\tau, \cos \theta_\tau, \sin \theta_\tau \right]. \tag{88}$$

Encoding heading by its sine and cosine avoids the discontinuity at the angular wrap-around boundary.

This integrator is a kinematic approximation rather than a complete vehicle-dynamics model. It constrains acceleration and local curvature but does not model tire slip, friction limits, suspension dynamics, or actuator dynamics. The formal result in Appendix A.6 therefore characterizes only the explicit bounds imposed by the discrete kinematic update and does not provide a guarantee of full dynamic feasibility.

B.3 Bounded Integration for Holonomic Agents

For a holonomic source such as a pedestrian, the physical anchor provides the initial planar position,

$$b_s = (x_0, y_0). \tag{89}$$

For each mode, the physical head predicts planar velocity logits

$$\widetilde{u}_{\tau,x}^{(k)}, \quad \widetilde{u}_{\tau,y}^{(k)}. \tag{90}$$

We bound the two components as

$$u_{\tau,x}^{(k)} = \frac{v_{\max}}{\sqrt{2}} \tanh \widetilde{u}_{\tau,x}^{(k)}, \quad u_{\tau,y}^{(k)} = \frac{v_{\max}}{\sqrt{2}} \tanh \widetilde{u}_{\tau,y}^{(k)}. \tag{91}$$

This parameterization ensures

$$\left\| \mathbf{u}_{\tau}^{(k)} \right\|_2 \leq v_{\max}. \tag{92}$$

Starting from b_s, the planar state evolves according to

$$\begin{aligned} x_{\tau+1} &= x_\tau + u_{\tau,x} \Delta t, \\ y_{\tau+1} &= y_\tau + u_{\tau,y} \Delta t. \end{aligned} \tag{93}$$

The corresponding mediator state is

$$n_\tau^{(k)} = \left[x_\tau, y_\tau, u_{\tau,x}, u_{\tau,y} \right]. \tag{94}$$

This branch does not impose vehicle-specific nonholonomic dynamics on pedestrian motion. Its role is to replace unconstrained coordinate regression with an explicit and bounded planar-motion representation.

B.4 Intrinsic Polyline Integration for Static Map Elements

For a static map source, the physical anchor contains the initial position and local tangent,

$$b_s = (x_0, y_0, \theta_0). \tag{95}$$

At segment τ, the physical head predicts a raw step-length logit $\widetilde{s}_\tau$ and two raw directional components $\widetilde{c}_\tau$ and $\widetilde{r}_\tau$. The positive step length is

$$\Delta s_\tau = \gamma \, \mathrm{Softplus}\left(\widetilde{s}_\tau \right). \tag{96}$$

The directional variables are normalized as

$$\begin{bmatrix} c_\tau \\ r_\tau \end{bmatrix} = \frac{\begin{bmatrix} \widetilde{c}_\tau \\ \widetilde{r}_\tau \end{bmatrix}}{\sqrt{\widetilde{c}_\tau^2 + \widetilde{r}_\tau^2 + \epsilon_{\mathrm{dir}}}}. \tag{97}$$

The heading increment is

$$\Delta \phi_\tau = \mathrm{atan2}(r_\tau, c_\tau), \tag{98}$$

and the geometric state evolves as

$$\begin{aligned} \theta_{\tau+1} &= \theta_\tau + \Delta \phi_\tau, \\ x_{\tau+1} &= x_\tau + \Delta s_\tau \cos \theta_{\tau+1}, \\ y_{\tau+1} &= y_\tau + \Delta s_\tau \sin \theta_{\tau+1}. \end{aligned} \tag{99}$$

The map mediator state is

$$n_\tau = \left[x_\tau, y_\tau, \cos \theta_\tau, \sin \theta_\tau, \Delta s_\tau \right]. \tag{100}$$

The intrinsic representation generates neighboring points through explicit local length and direction changes and therefore preserves sequential geometric continuity. It does not guarantee global map topology, absence of self-intersections, or semantic connectivity.

B.5 PHYSICAL-STATE RE-ENCODER

The mediator types contain different physical channels but are represented in the same ego-centric coordinate frame used by the host architecture and are eventually mapped to the common latent dimension d. To balance physical quantities with different units, we use fixed normalization statistics computed exclusively from the training split.

For continuous channel c of mediator type τ, define

$$s_{\tau,c} = \max\left\{Q_{0.95}\left(|g_{\tau,c}|\right), \epsilon_{\text{norm}}\right\}, \tag{101}$$

where $Q_{0.95}$ is the 95th percentile over all valid ground-truth states of that type in the training split and $\epsilon_{\text{norm}} = 10^{-6}$. For sine and cosine heading channels, we set $s_{\tau,c} = 1$. The fixed normalization operator is

$$\left[\mathcal{R}_\tau(n)\right]_c = \frac{n_c}{s_{\tau,c}}. \tag{102}$$

These statistics are computed before optimization, remain fixed throughout training, and are shared by all controlled variants. Validation and test data are never used to estimate them. The same normalization operator is used for the physical encoder and for the mediator-state supervision in Appendix C.

Given

$$N_s^{(k)} \in \mathbb{R}^{T \times d_{\text{phys}}}, \tag{103}$$

we first normalize the sequence and project it to the host latent dimension,

$$H_0 = \text{LN}\left(\mathcal{R}_{\tau_s}\left(N_s^{(k)}\right) W_{\text{in}} + b_{\text{in}}\right). \tag{104}$$

Temporal structure is encoded by two one-dimensional convolutional layers, each with width d, kernel size three, unit stride, and padding one:

$$H_\ell = \text{ReLU}\left(\text{Conv1D}_\ell(H_{\ell-1})\right), \quad \ell = 1, 2. \tag{105}$$

The temporal representation is aggregated as

$$z_s^{(k)} = W_{\text{out}}\left[\text{AvgPool}(H_2) \| \,\text{MaxPool}(H_2)\right]. \tag{106}$$

Vehicle-like, holonomic, and map mediators use encoders with the same architecture but separate parameters. The ego vehicle shares the vehicle-like encoder. The re-encoder does not fuse any semantic feature derived directly from the raw source query q_s. The mode probabilities remain part of the complete mediator representation and enter the subsequent interaction only as attention priors.

For attention head h,

$$K_{s,k}^h = W_K^h z_s^{(k)}, \quad V_{s,k}^h = W_V^h z_s^{(k)}. \tag{107}$$

B.6 PRIOR-WEIGHTED MODE-EXPANDED ATTENTION

All valid source-mode pairs are retained as independent attention tokens. For attention head h, the projected target query is

$$Q_t^h = W_Q^h q_t. \tag{108}$$

We compute the mode log-prior directly from the mode logits as

$$\lambda_s^{(k)} = \text{LogSoftmax}(\beta_s)_k = \log p_s^{(k)}. \tag{109}$$

The compatibility logit for source s and mode k is

$$e_{t,s,k}^h = \frac{(Q_t^h)^\top K_{s,k}^h}{\sqrt{d_h}} + \lambda_s^{(k)}. \tag{110}$$

The attention weights are normalized jointly over all valid sources and physical modes,

$$\alpha_{t,s,k}^h = \text{Softmax}_{s,k}\left(e_{t,s,k}^h\right), \tag{111}$$

and the output of one head is

$$m_t^h = \sum_{s \in \mathcal{S}_t} \sum_{k=1}^{K_s} \alpha_{t,s,k}^h V_{s,k}^h. \tag{112}$$

The multi-head output is

$$m_t = W_O \left[m_t^1 \| \cdots \| m_t^H \right]. \tag{113}$$

To make the probabilistic interpretation explicit, define the base measure over valid source-mode tokens as

$$\rho_t(s,k) = \frac{p_s^{(k)}}{|\mathcal{S}_t|}. \tag{114}$$

Since the probabilities of each source sum to one, ρ_t is a probability distribution. Equation 111 is equivalently the exponentially reweighted distribution

$$\alpha_{t,s,k}^h = \frac{\rho_t(s,k) \exp\left((Q_t^h)^\top K_{s,k}^h / \sqrt{d_h} \right)}{\sum_{s',j} \rho_t(s',j) \exp\left((Q_t^h)^\top K_{s',j}^h / \sqrt{d_h} \right)}. \tag{115}$$

Consequently,

$$m_t^h = \mathbb{E}_{(S,K) \sim \alpha_t^h} \left[V_{S,K}^h \right]. \tag{116}$$

The message is therefore an exact expectation under a query-conditioned discrete distribution over physically grounded source-mode tokens. The predicted mode probabilities provide the prior mass, while target-to-mediator compatibility determines the target-dependent reweighting.

The target representation is updated using the residual convention of the host architecture,

$$q_t' = q_t + m_t. \tag{117}$$

Layer normalization and the feed-forward block retain the ordering of the host model so that CIB does not introduce an additional Transformer redesign.

B.7 Source-Side No-Bypass Implementation

The structural mediation property requires an implementation-level constraint on every source-dependent input to a CIB-replaced interaction edge. The source descriptor

$$\chi_s = (q_s, b_s, \tau_s) \tag{118}$$

may influence that interaction only through

$$\chi_s \longrightarrow \mathcal{M}_s = \left(\tau_s, \left\{ N_s^{(k)}, p_s^{(k)} \right\}_k \right) \longrightarrow q_t'. \tag{119}$$

Accordingly, the raw source query is not used as an attention key or value. A raw source-anchor embedding is not added separately to the physical key, value, or attention logit. Source classification confidence, quality scores, and other source-specific semantic features are likewise excluded from the CIB attention bias. When the host block uses relative-position information, the source-side geometric component is recomputed from the corresponding physical hypothesis $N_s^{(k)}$ rather than obtained directly from the raw anchor b_s.

The active source set and validity mask are determined by the host model before the CIB interaction and are treated as fixed conditioning variables in Proposition 1. The mask only determines whether a source-mode token is valid and does not inject an additional continuous source-dependent score into the attention logits.

The target query, target anchor, and target residual state remain available to the interaction. These quantities do not constitute a source-side bypass because structural mediation constrains information newly transmitted from the source rather than information already stored in the target.

Temporal self-interaction, identity propagation, and perception operations outside a replaced cross-entity edge are retained from the host architecture. The structural guarantee therefore applies specifically to the CIB-mediated interaction channel and does not assert conditional independence of the complete host network from the original source representation.

B.8 Initialization and Numerical Stability

The generator trunk, physical encoders, and attention projections use the same initialization convention as the corresponding host-network layers. Learned mode embeddings are initialized independently using the same rule as the host query embeddings.

The final physical-parameter projections use a near-zero initialization,

$$W_{\mathrm{ctrl}} \sim \mathcal{N}\left(0, \sigma_{\mathrm{ctrl}}^2\right), \quad \sigma_{\mathrm{ctrl}} = 10^{-3}, \tag{120}$$

with zero bias. For the vehicle branch, the initial controls therefore satisfy

$$a_\tau \simeq 0, \quad \kappa_\tau \simeq 0, \tag{121}$$

so that the initial rollout remains close to a low-curvature continuation from the current physical state. Independent mode embeddings and small nonzero control weights prevent the multimodal hypotheses from being exactly identical at initialization.

The mode-scoring head is initialized with zero weights and zero bias, yielding the uniform initial mode distribution

$$p_s^{(k)} = \frac{1}{K_s}. \tag{122}$$

For the map branch, the initial direction follows the tangent of the map anchor. The cosine-direction bias is initialized to one and the sine-direction bias to zero. We use

$$\gamma = 1.0\ \mathrm{m}, \quad b_{\mathrm{step}} = \mathrm{Softplus}^{-1}(1) = \log(e-1), \tag{123}$$

which initializes the step length to approximately one meter.

Mode log-priors are computed directly with `log_softmax`, avoiding an explicit probability clamp. Direction normalization and physical-state normalization use numerical floors of $\epsilon_{\mathrm{dir}} = 10^{-6}$ and $\epsilon_{\mathrm{norm}} = 10^{-6}$, respectively.

These choices provide a stable initial physical state and break exact permutation symmetry among modes. They do not constitute a convergence guarantee or a bound on gradient magnitude.

C Training Objectives and Optimization

CIB preserves the original perception, motion-prediction, and planning objectives of the host model and adds explicit supervision on the physical mediators that participate in cross-entity interaction. This auxiliary objective is not a causal-effect estimator and is not treated as a separate causal loss. Its purpose is to ensure that the physical hypotheses contained in $\mathcal{M}_s$ retain well-defined kinematic or geometric semantics.

Supervision is applied to the integrated physical states rather than directly to acceleration, curvature, or other internal control variables. Both the downstream task losses and the mediator losses can therefore propagate through the differentiable integrators to the physical-parameter predictor.

C.1 Overall Training Objective

Let $\mathcal{L}_{\mathrm{host}}$ denote the original training objective of the host architecture, including its perception, motion-prediction, and planning losses. The complete objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{host}} + \lambda_{\mathrm{phys}} \mathcal{L}_{\mathrm{phys}}. \tag{124}$$

Let Ω^+ denote the set of CIB channel-stage pairs that contain valid mediator supervision in the current minibatch. For $(c, r) \in \Omega^+$, let $\mathcal{L}_{\mathrm{med}}^{c,r}$ denote the corresponding mediator loss after normalization over valid entities, future states, and supervised state channels. We define

$$\mathcal{L}_{\mathrm{phys}} = \frac{1}{|\Omega^+|} \sum_{(c,r)\in\Omega^+} \mathcal{L}_{\mathrm{med}}^{c,r}. \tag{125}$$

A channel-stage pair with no valid matched entity or ground-truth state in the current minibatch is excluded from Ω^+ rather than included as a zero-valued term.

Dynamic sources, including the ego vehicle, vehicle-like agents, and holonomic agents, use the dynamic-mediator objective defined below. Static map sources use the map-mediator objective. If a CIB block contains multiple source types, we first normalize each type-specific loss independently and then average over the source types that are present in the minibatch.

This construction controls the aggregate scale of auxiliary supervision across architectures with different numbers of CIB blocks. It does not imply equal gradient norms across tasks or interaction paradigms.

C.2 Supervision of Multimodal Dynamic Mediators

Whenever the host perception or motion module already provides a query-to-instance assignment, CIB reuses that assignment and does not redefine the instance identity. If a particular host stage provides no assignment, we use the same matching procedure as the corresponding host task (Kuhn, 1955; Carion et al., 2020).

For a matched dynamic entity i, CIB predicts

$$\left\{N_i^{(k)}, p_i^{(k)}\right\}_{k=1}^{K_i}. \tag{126}$$

Let $m_{i,\tau} \in \{0,1\}$ indicate whether the ground-truth future position at step τ is valid. We select the supervised mode using average displacement in position space,

$$d_i^{(k)} = \frac{\sum_\tau m_{i,\tau} \left\| \widehat{\mathbf{p}}_{i,\tau}^{(k)} - \mathbf{p}_{i,\tau}^{\mathrm{gt}} \right\|_2}{\sum_\tau m_{i,\tau}}, \tag{127}$$

and

$$k_i^\star = \arg\min_k d_i^{(k)}. \tag{128}$$

Entities with no valid future state are excluded. The discrete Equation 128 is used only for assignment and receives no gradient. Selecting the winner from positional displacement keeps the mode assignment independent of the relative weighting of auxiliary speed and heading channels, and follows the standard multiple-choice formulation used in multimodal trajectory prediction (Chai et al., 2020).

The mode probabilities are supervised with

$$\mathcal{L}_{\mathrm{mode}} = -\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \log p_i^{(k_i^\star)}, \tag{129}$$

where $\mathcal{I}$ contains the matched entities with at least one valid future state.

Ground-truth physical states are constructed in the same coordinate frame as the source anchors. Positions are taken directly from the benchmark annotations. When a reliable heading annotation is available, the corresponding yaw is used directly. Otherwise, the heading channels are masked from the auxiliary state loss rather than reconstructed from unstable finite differences at low speed. Velocity channels are obtained from consecutive valid positions:

$$\mathbf{u}_{i,\tau}^{\mathrm{gt}} = \frac{\mathbf{p}_{i,\tau}^{\mathrm{gt}} - \mathbf{p}_{i,\tau-1}^{\mathrm{gt}}}{\Delta t}, \tag{130}$$

where the current source anchor supplies $\mathbf{p}_{i,0}^{\mathrm{gt}}$. Vehicle speed is the Euclidean norm of this velocity, while holonomic agents retain its planar components.

Let $a_{i,\tau,c} \in \{0,1\}$ indicate whether state channel c is supervised for entity i at time τ. Using the fixed normalization from Equation 102, the state loss is

$$\mathcal{L}_{\mathrm{state}} = \frac{\sum_{i,\tau,c} m_{i,\tau} a_{i,\tau,c} \left| \left[\mathcal{R}_{\tau_i}\left(\widehat{n}_{i,\tau}^{(k_i^\star)}\right)\right]_c - \left[\mathcal{R}_{\tau_i}\left(n_{i,\tau}^{\mathrm{gt}}\right)\right]_c \right|}{\sum_{i,\tau,c} m_{i,\tau} a_{i,\tau,c}}. \tag{131}$$

The complete dynamic-mediator objective is

$$\mathcal{L}_{\text{dyn}} = \mathcal{L}_{\text{state}} + \lambda_{\text{mode}}\mathcal{L}_{\text{mode}}. \tag{132}$$

Acceleration, curvature, and steering angle are not directly supervised. Their gradients are obtained through the differentiable physical rollout. We do not introduce an additional entropy or diversity objective. Mode symmetry is broken by the learned mode embeddings and the independent near-zero initialization described in Appendix B.8.

C.3 Supervision of Static Map Mediators

Static map queries reuse both the query-to-instance assignment and the ground-truth point permutation selected by the host mapping head. This keeps the CIB geometry supervision consistent with the host map-regression target and avoids introducing a second orientation convention for geometrically equivalent polylines.

Let

$$\left\{\mathbf{p}_\tau^{\text{gt}}\right\}_{\tau=0}^{T_m-1} \tag{133}$$

denote the ordered ground-truth polyline after host assignment. For $\tau = 0, \ldots, T_m - 2$, define

$$\mathbf{d}_\tau = \mathbf{p}_{\tau+1}^{\text{gt}} - \mathbf{p}_\tau^{\text{gt}}, \quad \Delta s_\tau^{\text{gt}} = \|\mathbf{d}_\tau\|_2 . \tag{134}$$

For nondegenerate segments, the ground-truth tangent is

$$\theta_{\tau+1}^{\text{gt}} = \operatorname{atan2}\left(d_{\tau,y}, d_{\tau,x}\right), \tag{135}$$

and the corresponding physical state is

$$n_{\tau+1}^{\text{gt}} = \left[x_{\tau+1}, y_{\tau+1}, \cos\theta_{\tau+1}, \sin\theta_{\tau+1}, \Delta s_\tau\right]. \tag{136}$$

If

$$\Delta s_\tau < \epsilon_{\text{seg}}, \quad \epsilon_{\text{seg}} = 10^{-3}\ \text{m}, \tag{137}$$

the segment is treated as geometrically degenerate. Its direction channels are excluded from the auxiliary map-state loss because the local tangent is not well defined, while the point-position channels remain supervised. The threshold is used only as a numerical tolerance and is fixed across all datasets and controlled variants.

Let $a_{i,\tau,c}^{\text{map}}$ denote the resulting channel-validity mask. The normalized map-mediator loss is

$$\mathcal{L}_{\text{map}} = \frac{\displaystyle\sum_{i,\tau,c} a_{i,\tau,c}^{\text{map}} \left| \left[\mathcal{R}_{\text{map}}\left(\widehat{n}_{i,\tau}\right)\right]_c - \left[\mathcal{R}_{\text{map}}\left(n_{i,\tau}^{\text{gt}}\right)\right]_c \right|}{\displaystyle\sum_{i,\tau,c} a_{i,\tau,c}^{\text{map}}}. \tag{138}$$

Positive segment lengths and recursive geometric continuity are already enforced by the forward parameterization. We therefore do not add an auxiliary smoothness loss that would duplicate the same structural prior. Map classification, instance existence, and global localization remain supervised by the original mapping objective in $\mathcal{L}_{\text{host}}$.

C.4 Temporal Alignment for Iterative Refinement

Iterative architectures may generate mediator predictions at multiple refinement stages. We align every prediction to absolute future time before computing either mode assignment or state supervision.

Let $r = 1, \ldots, R$ index the refinement stage and let o_r denote the number of prediction intervals by which physical time has advanced relative to the original observation. For prediction index $j = 1, \ldots, T_{\text{pred}}$ at stage r, the corresponding ground-truth state is

$$n_{o_r+j}^{\text{gt}}. \tag{139}$$

Table 4: **Training schedule for controlled SparseDrive experiments.** Paired baselines use the same Stage 2 optimization budget as the corresponding CIB model.

Variant	Stage	Batch	Epochs	LR	Weight Decay	λ_{phys}
Sequential	1	64	100	4×10^{-4}	10^{-3}	0
	2	48	10	3×10^{-4}	10^{-3}	0.2
Parallel	1	64	100	4×10^{-4}	10^{-3}	0
	2	48	10	3×10^{-4}	10^{-3}	0.2
Iterative	1	64	100	4×10^{-4}	10^{-3}	0
	2	32	10	2×10^{-4}	10^{-3}	0.2

Using one-based indexing, the validity mask is

$$m_{r,j} = \mathbb{1}\left[o_r + j \le T_{\text{gt}}\right]. \tag{140}$$

The same mask is used for mode assignment and physical-state regression. Predictions outside the annotated future support contribute no auxiliary gradient.

In the PhysDrive-I implementation, each recurrent planning step advances one prediction interval, so

$$o_r = r - 1. \tag{141}$$

This alignment depends only on the rollout schedule and does not interpret the refinement index as a causal intervention.

C.5 Loss Aggregation across Interaction Channels

Each dynamic or map mediator loss is first normalized over its own valid entities, future states, and supervised state channels. The normalized channel-stage losses are then averaged over Ω^+ according to Equation 125.

The mode-classification coefficient λ_{mode} is applied only inside each dynamic-mediator objective. The global coefficient λ_{phys} is applied once to the aggregated physical objective in Equation 124; it is not applied separately to every CIB block.

This averaging makes the first-order coefficient scale of physical supervision insensitive to the number of active CIB instances. It should not be interpreted as a gradient-balancing method. In particular, equal aggregate loss coefficients do not imply equal values of $\|\nabla_\theta \mathcal{L}\|$ across architectures, tasks, or training stages.

C.6 Optimization Protocol and Hyperparameters

We use a two-stage optimization protocol following the SparseDrive host training procedure (Sun et al., 2025). Stage 1 trains the host perception representation. Stage 2 starts from the corresponding Stage 1 checkpoint and introduces the motion-planning head together with the CIB modules. The stage-wise optimization schedule is summarized in Table 4, while the default CIB hyperparameters are listed in Table 5.

Within each controlled comparison and random seed, the host baseline, DSB, and CIB variants are initialized from the same Stage 1 checkpoint. The optimizer and learning-rate scheduler are reinitialized at the beginning of Stage 2. All model parameters remain trainable, with a learning-rate multiplier of 0.1 applied to the image backbone during Stage 2.

We use AdamW (Loshchilov & Hutter, 2017) with weight decay 10^{-3}. A 500-iteration linear warmup with warmup ratio $1/3$ is followed by cosine learning-rate decay with minimum learning-rate ratio 10^{-3}. Gradients are clipped by their global ℓ_2 norm with maximum norm 25. Mixed-precision training follows the host implementation.

For a paired host baseline, Stage 2 uses the same epochs, batch size, learning rate, scheduler, and checkpoint initialization but omits $\mathcal{L}_{\text{phys}}$. DSB uses the same Stage 2 protocol and mediator supervision as CIB. Consequently, differences in the controlled study cannot be attributed to an additional fine-tuning budget.

For the three-seed matched-control experiment, each seed independently runs both training stages rather than varying only the Stage 2 initialization.

Table 5: **Default CIB hyperparameters.** Temporal horizons and representation dimensions follow the SparseDrive host configuration. Physical bounds are fixed across controlled comparisons. We additionally use the fixed numerical tolerance $\epsilon_{\text{seg}} = 10^{-3}$ m for degenerate map segments.

Parameter	Value	Parameter	Value
Dynamic modes K	6	Map modes K_{map}	1
Agent horizon	12	Ego horizon	6
Time interval Δt	0.5 s	Map samples	20
Latent dimension d	256	Attention heads H	8
Vehicle $a_{\max}$	6.0 m/s^2	Vehicle $\kappa_{\max}$	0.20 m^{-1}
Holonomic $v_{\max}$	4.0 m/s	Map scale γ	1.0 m
Control init. σ_{ctrl}	10^{-3}	Mode weight λ_{mode}	0.1
Physical weight λ_{phys}	0.2	Attention dropout	0.1
ϵ_{dir}	10^{-6}	ϵ_{norm}	10^{-6}

D Evaluation Protocols and Metric Definitions

This section specifies the evaluation protocols and diagnostic metrics used throughout the paper. We distinguish standard benchmark performance from controlled robustness evaluation and mechanism-oriented diagnostics. Within each controlled comparison, all models use the same data split, prediction horizon, evaluator, and post-processing procedure. Results quoted directly from prior work without re-evaluation under our protocol are explicitly marked as reported results and are not used to establish strict relative improvements.

D.1 Benchmark Versions and Reporting Policy

We distinguish results reproduced under our frozen evaluation stack from results quoted directly from prior work. Only the former are used to establish strict paired improvements between methods.

All controlled nuScenes experiments use the same validation split and planning evaluator. The primary NAVSIM comparison follows the v1.1 protocol and reports PDMS (Dauner et al., 2024). Closed-loop evaluation uses Bench2Drive v0.0.4 with CARLA 0.9.15 and the official 220-route validation set (Jia et al., 2024). Failed and crashed routes remain part of the final evaluation.

D.2 nuScenes Open-Loop Planning Metrics

We use a single SparseDrive-style planning evaluator for all controlled nuScenes comparisons. Let $\widehat{\mathbf{p}}_{i,j}$ and $\mathbf{p}^{\text{gt}}_{i,j}$ denote the predicted and ground-truth ego positions for sample i at future index j, respectively. The temporal interval is $\Delta t = 0.5$ s. For an evaluation horizon $H \in \{1, 2, 3\}$ seconds, define

$$\mathcal{J}_H = \{j : j\Delta t \le H\}\,. \tag{142}$$

The prefix-averaged displacement error is

$$\text{L2}(H) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{J}_H|}\sum_{j\in\mathcal{J}_H}\left\|\widehat{\mathbf{p}}_{i,j} - \mathbf{p}^{\text{gt}}_{i,j}\right\|_2. \tag{143}$$

Hence, $\text{L2}(2\,\text{s})$ averages all prediction errors from 0.5 s through 2.0 s rather than evaluating only the endpoint at 2.0 s. We state this convention explicitly because earlier end-to-end driving work has used both endpoint and prefix-averaged reporting conventions (Hu et al., 2023; Jiang et al., 2023; Sun et al., 2025).

Let $c_{i,j} \in \{0, 1\}$ be the collision indicator returned by the common SparseDrive collision evaluator at future index j. We report

$$\text{CR}(H) = \frac{100}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{J}_H|}\sum_{j\in\mathcal{J}_H} c_{i,j}. \tag{144}$$

This quantity is the average of per-timestep collision indicators over the prefix and should not be interpreted as the cumulative event probability

$$P\left(\exists j \le H : c_j = 1\right). \tag{145}$$

Consequently, $\mathrm{CR}(1\,\mathrm{s})$, $\mathrm{CR}(2\,\mathrm{s})$, and $\mathrm{CR}(3\,\mathrm{s})$ are not mathematically required to be monotonic.

The aggregate metrics are

$$\mathrm{L2}_{\mathrm{Avg}} = \frac{1}{3} \sum_{H \in \{1,2,3\}} \mathrm{L2}(H) \tag{146}$$

and

$$\mathrm{CR}_{\mathrm{Avg}} = \frac{1}{3} \sum_{H \in \{1,2,3\}} \mathrm{CR}(H). \tag{147}$$

All CIB models, matched bottlenecks, and host baselines in our controlled tables are evaluated using exactly this protocol. Numbers taken directly from prior publications are reported separately when the original checkpoint cannot be evaluated under the same implementation.

D.3 NAVSIM Non-Reactive Evaluation

Our primary NAVSIM comparison uses the NAVSIM v1.1 PDMS protocol (Dauner et al., 2024). NAVSIM v1 performs a four-second non-reactive pseudo-simulation in which background traffic follows recorded trajectories while the predicted ego trajectory is simulated and scored.

For one scene, the Predictive Driver Model Score is

$$\mathrm{PDMS} = \mathrm{NC} \cdot \mathrm{DAC} \cdot \frac{5\,\mathrm{TTC} + 5\,\mathrm{EP} + 2\,\mathrm{C}}{12}, \tag{148}$$

where NC denotes no-at-fault collision, DAC denotes drivable-area compliance, TTC denotes time-to-collision compliance, EP denotes ego progress, and C denotes comfort. We report the mean scene-level PDMS over the evaluation split.

D.4 Bench2Drive Closed-Loop Evaluation

All new closed-loop evaluations use Bench2Drive v0.0.4 with CARLA 0.9.15. The official validation set contains 220 routes covering 44 scenarios with five routes per scenario. Failed and crashed routes remain part of the evaluation set.

For route i, let $r_i \in [0,1]$ denote its route-completion fraction and let $p_{i,j} \in [0,1]$ denote its j-th infraction penalty. Following the official benchmark, Driving Score is

$$\mathrm{DS} = \frac{100}{N} \sum_{i=1}^{N} r_i \prod_j p_{i,j}. \tag{149}$$

Success Rate is

$$\mathrm{SR} = \frac{100 N_{\mathrm{success}}}{N}. \tag{150}$$

Route success is taken directly from the official evaluator rather than redefined by our method. Driving efficiency, comfort, and multi-ability scores are reported as additional diagnostics when applicable, while DS and SR serve as the primary closed-loop metrics.

D.5 Metrics for Controlled Interventions

Controlled-intervention metrics quantify behavioral differences between paired scenes in which a specified factor is changed. They are diagnostic sensitivity measures rather than estimators of a causal effect.

Let $\widehat{\mathbf{p}}_{i,j}^{\mathrm{obs}}$ and $\widehat{\mathbf{p}}_{i,j}^{\mathrm{int}}$ denote the observational and intervened plans. Reaction Sensitivity is

$$\mathrm{RS}(H) = \frac{1}{N_{\mathrm{int}}} \sum_i \frac{1}{|\mathcal{J}_H|} \sum_{j \in \mathcal{J}_H} \left\| \widehat{\mathbf{p}}_{i,j}^{\mathrm{obs}} - \widehat{\mathbf{p}}_{i,j}^{\mathrm{int}} \right\|_2. \tag{151}$$

RS measures response magnitude and has no universal monotonic interpretation. A small response is desirable for behaviorally irrelevant nuisance changes, whereas an excessively small response to a physically relevant agent change may indicate insensitivity to the traffic interaction.

For appearance-only pairs, we define

$$\mathrm{NIS}(H) = \frac{1}{N_{\mathrm{pair}}} \sum_i \frac{1}{|\mathcal{J}_H|} \sum_{j \in \mathcal{J}_H} \left\| \widehat{\mathbf{p}}_{i,j}^{z} - \widehat{\mathbf{p}}_{i,j}^{z'} \right\|_2 . \tag{152}$$

The paired scenes share the same ego state, dynamic-agent states, map geometry, route command, and temporal indexing and differ only in the controlled rendering condition. Lower NIS therefore indicates greater stability to the specific appearance changes considered here, not invariance to arbitrary domain shift.

False Braking Rate is evaluated only on the eligible subset $\mathcal{I}_{\mathrm{free}}$ for which the intervention removes the primary conflict and geometric collision checking confirms that strong braking is not required.

For predicted ego waypoints, we compute the finite-difference speed and longitudinal acceleration as

$$v_{i,j} = \frac{\|\widehat{\mathbf{p}}_{i,j} - \widehat{\mathbf{p}}_{i,j-1}\|_2}{\Delta t}, \quad a_{i,j} = \frac{v_{i,j} - v_{i,j-1}}{\Delta t}. \tag{153}$$

We define the fixed diagnostic braking threshold as

$$a_{\mathrm{FB}} = 2.0 \ \mathrm{m/s^2}. \tag{154}$$

False Braking Rate is then

$$\mathrm{FBR} = \frac{100}{|\mathcal{I}_{\mathrm{free}}|} \sum_{i \in \mathcal{I}_{\mathrm{free}}} \mathbf{1}\left[\min_j a_{i,j} < -a_{\mathrm{FB}} \right]. \tag{155}$$

The threshold is fixed before final model evaluation and is used only to identify pronounced unnecessary deceleration. It is not interpreted as a universal comfort or safety boundary.

D.6 Iterative Consistency and Ride Comfort

For an iterative model, let $\widehat{\mathbf{p}}_i^{(r)}(t_i + H)$ denote the position predicted at refinement stage r for the absolute future time $t_i + H$. Predictions are first aligned in absolute time using the procedure in Appendix C.4. Relative to the first refinement stage, Step Divergence is

$$\mathrm{SD}^{(r)}(H) = \frac{1}{N} \sum_{i=1}^{N} \left\| \widehat{\mathbf{p}}_i^{(r)}(t_i + H) - \widehat{\mathbf{p}}_i^{(1)}(t_i + H) \right\|_2 . \tag{156}$$

The main diagnostic table reports $\mathrm{SD}^{(R)}(H)$ at 1, 2, and 3 seconds for the final refinement stage R.

A lower Step Divergence indicates greater refinement stability but does not by itself imply higher planning accuracy. We therefore report the standard planning error at each refinement stage together with SD to rule out the degenerate case in which a model appears stable simply because it performs little refinement.

Ride comfort is summarized by mean absolute jerk. Using the finite-difference velocity and acceleration defined above,

$$j_{i,j} = \frac{a_{i,j} - a_{i,j-1}}{\Delta t}, \tag{157}$$

and

$$\mathrm{MAJ} = \frac{1}{\sum_i (T_i - 2)} \sum_i \sum_{j=3}^{T_i} |j_{i,j}| . \tag{158}$$

We use absolute jerk rather than signed average jerk so that positive and negative changes in acceleration cannot cancel.

D.7 Maneuver Subsets and Geometric Smoothness

For maneuver-stratified analysis, we construct subsets directly from the ground-truth ego trajectory. Let

$$\Delta\psi_i = \mathrm{wrap}\left(\psi_{i,T} - \psi_{i,1}\right). \tag{159}$$

A scene is classified as straight when $|\Delta\psi_i| \leq 5^\circ$ and as turning when $|\Delta\psi_i| \geq 15^\circ$. Intermediate cases are excluded. This partition is used to analyze performance across maneuver geometries and is not interpreted as an out-of-distribution generalization benchmark.

For the map analysis, let

$$\ell_j = \|\mathbf{p}_{j+1} - \mathbf{p}_j\|_2 \tag{160}$$

and

$$\phi_j = \operatorname{atan2}\left(y_{j+1} - y_j, x_{j+1} - x_j\right). \tag{161}$$

We define the discrete curvature

$$\kappa_j = \frac{\operatorname{wrap}(\phi_j - \phi_{j-1})}{(\ell_j + \ell_{j-1})/2} \tag{162}$$

and Curvature Variation per Meter

$$\mathrm{CVM} = \frac{\sum_j |\kappa_{j+1} - \kappa_j|}{\sum_j (\ell_j + \ell_{j+1})/2}. \tag{163}$$

CVM has units of m^{-2} and measures local curvature smoothness. It does not measure semantic map correctness, topological validity, or global localization accuracy and is therefore interpreted together with the standard mapping metrics.

D.8 Statistical Reporting

For the central matched-control experiment, each model is trained with three independent random seeds and reported using mean and standard deviation. These statistics quantify optimization variability across independently trained models.

Cross-paradigm experiments conducted with one trained checkpoint are explicitly identified as single-training-run evaluations and are not presented as estimates of training variance.

For Bench2Drive, we additionally quantify route-level evaluation uncertainty using a paired nonparametric bootstrap over the official 220 validation routes. For a comparison between methods A and B, one bootstrap replicate samples 220 route indices with replacement and applies the same sampled index set to both methods. We then compute the paired differences

$$\Delta\mathrm{DS}^{(b)} = \mathrm{DS}_B^{(b)} - \mathrm{DS}_A^{(b)}, \quad \Delta\mathrm{SR}^{(b)} = \mathrm{SR}_B^{(b)} - \mathrm{SR}_A^{(b)}. \tag{164}$$

We use 10,000 bootstrap replicates and report the 2.5th and 97.5th percentiles as the 95% confidence interval.

This bootstrap quantifies uncertainty induced by the finite evaluation-route set for a fixed trained checkpoint. It is distinct from training-seed variability and should not be interpreted as a confidence interval over model training.

Controlled-intervention metrics use paired resampling for the same reason: the observational and intervened versions of a scene remain paired in every bootstrap replicate.

E Mechanism Validation and Controlled Interventions

This section evaluates the mechanism through four complementary controls: a matched Direct-State Bottleneck, frozen representation probing, appearance-only paired interventions, and physically relevant agent interventions. The four analyses address distinct alternatives to the proposed interpretation: generic bottleneck regularization, residual appearance decodability, sensitivity to controlled visual changes, and sensitivity to changes in the physical interaction state.

Unless otherwise specified, all mechanism studies use the same SparseDrive host configuration, optimization schedule, number of physical modes, and attention structure as the corresponding controlled comparison. Experimental thresholds and intervention rules are fixed before evaluating the final models.

E.1 Matched Direct-State Bottleneck

To separate explicit physical parameterization from generic bottleneck regularization, we construct a Direct-State Bottleneck, denoted as DSB. DSB retains the CIB generator trunk, mode embeddings, mode-probability head, intermediate state supervision, physical-state re-encoder, prior-weighted mode-expanded attention, and optimization schedule. Its defining difference is that the mediator state is predicted directly rather than generated through a kinematic or geometric integrator.

For source s and mode k,

$$\widehat{N}_s^{(k)} = \Psi_{\tau_s}\left(h_s^{(k)}\right). \tag{165}$$

The predicted state channels follow the same target representation and normalization used by CIB, but no curvature, bounded planar-motion, or intrinsic-polyline integration is applied.

Because the direct-state and physical-control heads have different output dimensions, exact layer-wise parameter identity is neither possible nor necessary. We select the DSB hidden width before training using parameter count only, without consulting validation performance, so that its trainable capacity remains close to that of the corresponding CIB model.

The resulting parameter count, FLOPs, memory footprint, and throughput are reported in Appendix F.7. DSB and CIB use the same Stage 2 optimization budget, mediator-state supervision, number of modes, physical-state encoder, and downstream interaction structure. The comparison therefore controls for the presence of an intermediate supervised bottleneck while removing explicit physical integration.

E.2 Probing Controlled Appearance Information

The structural results in Appendix A constrain how source-side information enters a CIB interaction but do not imply that contextual information is absent from the mediator. We therefore measure the linear decodability of controlled appearance attributes from several internal representations.

After training the driving model, all of its parameters are frozen. For source s, we probe the raw latent query q_s and the mode-prior vector

$$\boldsymbol{\pi}_s = \left[p_s^{(1)}, \ldots, p_s^{(K_s)}\right]. \tag{166}$$

We additionally construct the probability-weighted physical summary

$$\overline{N}_s = \sum_{k=1}^{K_s} p_s^{(k)} N_s^{(k)} \tag{167}$$

and the probability-weighted encoded summary

$$\overline{z}_s = \sum_{k=1}^{K_s} p_s^{(k)} z_s^{(k)}. \tag{168}$$

These summaries are used only for diagnostic probing and are not part of the CIB forward interaction.

Because representation dimension can itself affect probe capacity, we include a fixed random projection of q_s matched to the dimension of each compared probe representation. Random projections are sampled once and remain frozen.

For appearance classification, source representations of the same type are pooled within each scene so that each rendered scene contributes one probe example. We train an ℓ_2-regularized multinomial linear classifier while keeping the driving model frozen. The four appearance conditions are balanced, and we report balanced accuracy and macro-F1. All renderings derived from the same underlying physical scene are assigned to the same data split, preventing paired views from leaking across probe training and evaluation.

Probe accuracy is interpreted only as linear decodability. It is not an estimate of mutual information and does not establish statistical independence. In particular, the explicit probe on $\boldsymbol{\pi}_s$ tests

Table 6: **Controlled appearance conditions used for paired CARLA interventions.** All conditions share the same physical actor states and navigation input.

Condition	Cloudiness	Precipitation	Wetness	Fog Density
Clear	0	0	0	0
Cloudy	80	0	0	0
Rain	80	60	80	0
Fog	20	0	0	40

whether the mode-probability pathway retains appearance information that may not be visible from an averaged physical-state summary.

As a complementary test, we train a linear ridge regressor to recover task-relevant physical quantities such as relative position, speed, and heading. This analysis examines whether reduced appearance decodability is accompanied by retention of physically relevant information.

E.3 Appearance-Only Paired Interventions

We construct paired appearance interventions using CARLA (Dosovitskiy et al., 2017). CARLA `WeatherParameters` modify RGB-camera rendering without changing actor physics, allowing the visual condition to vary while the underlying physical scene is held fixed.

We use Bench2Drive v0.0.4 with CARLA 0.9.15. For each selected route, we record the complete temporal input window required by the host model, including ego state, dynamic-actor transforms, traffic-light state, navigation command, and sensor calibration. The recorded physical sequence is then replayed under the appearance conditions listed in Table 6.

During replay, actor transforms and traffic-light states are restored at every input timestep, and the sensor extrinsics, route command, and temporal indexing remain unchanged. Weather parameters are therefore the intended source of variation across each paired rendering.

Clear weather serves as the reference condition. We report NIS separately for each appearance transformation and as an average across the controlled conditions. Lower NIS indicates greater stability to these specific rendering changes. The experiment does not establish invariance to arbitrary real-world weather, sensor shift, or geographic domain shift.

For representation probing, paired renderings derived from one physical scene always remain in the same data split. Probe models are selected without using the held-out intervention evaluation set.

E.4 Physically Relevant Agent Interventions

Reduced appearance sensitivity alone is insufficient to establish desirable behavior. A planner that ignores its surroundings could also achieve a low nuisance sensitivity. We therefore modify the physical state of a behaviorally relevant traffic participant and test whether the planner remains responsive to changes in the interaction state.

For Bench2Drive routes with an explicitly designated dynamic scenario actor, we use the actor specified by the benchmark metadata. When no unique actor is available, we select the dynamic actor within 30 m ahead of the ego vehicle with the smallest positive constant-velocity time to collision. Actor selection is completed before model inference and is shared by all compared methods.

We first perform an *agent-removal intervention*. The selected actor is absent throughout the complete temporal input window, while the ego state, all remaining actors, the map, the navigation command, and the rendering condition remain unchanged. Removing the actor throughout the temporal window avoids the history inconsistency introduced by masking only its current latent representation.

We additionally construct *speed interventions* for lane-associated wheeled actors. Let v_{obs} denote the observational longitudinal speed. We set

$$v_{\mathrm{int}} = \mathrm{clip}\left(v_{\mathrm{obs}} + \Delta v, 0, 15\right), \tag{169}$$

where

$$\Delta v \in \{-2, +2\}\ \mathrm{m/s}. \tag{170}$$

Table 7: **Paired route-level bootstrap on Bench2Drive.** Intervals are 95% percentile confidence intervals over the official 220-route validation set for fixed trained checkpoints. They quantify route-set uncertainty and do not include training-seed variability.

Comparison	ΔDS	ΔSR (pp)
PhysDrive-S − SparseDrive-S	3.0 [1.1, 4.9]	2.3 [−1.1, 5.7]
PhysDrive-P − SparseDrive-P	3.4 [1.4, 5.4]	2.7 [−0.9, 6.3]
PhysDrive-I − SparseDrive-I	3.4 [1.5, 5.2]	1.8 [−1.6, 5.2]

Table 8: **Representation probing.** Appearance prediction contains four balanced classes and therefore has a chance accuracy of 25%. Physical-state nMAE evaluates the retention of task-relevant geometric and kinematic information.

Representation	Appearance BA (%)	Macro-F1 (%)	Physical nMAE ↓
Raw query q_s	88.6	88.3	0.112
Random projection Rq_s	79.4	79.0	0.126
DSB state	66.8	66.0	0.095
Physical summary $\overline{N}_s$	**55.8**	**55.1**	**0.081**
Encoded summary $\overline{z}_s$	59.6	59.0	0.084
Mode prior $\boldsymbol{\pi}_s$	56.6	55.7	–

Starting from the first frame of the temporal input window, the intervened actor is replayed along its observational lane path with the modified longitudinal speed while preserving lane association and heading continuity. Pedestrians, crossing cyclists, and other actors for which lane-following longitudinal motion is not well defined are excluded from the speed intervention but can still participate in the removal experiment.

We quantify the trajectory change using Reaction Sensitivity from Appendix D.5. Reaction Sensitivity is reported without a preferred monotonic direction because a larger response is not universally better. For agent removal, False Braking Rate is evaluated only on samples for which geometric collision checking confirms that removing the selected actor eliminates the primary conflict.

The appearance and physical interventions are interpreted jointly. A model that suppresses all input sensitivity may achieve low NIS while failing to respond to a relevant change in the traffic state. The behavior targeted by CIB is therefore *selective sensitivity*: reduced response to controlled appearance variation while retaining a meaningful response to physically relevant interaction changes.

E.5 Latent Perturbation Robustness

We retain latent Gaussian perturbation as a secondary stress test of source representation corruption. It is not treated as a simulator intervention or as evidence of causal identification.

For source query q_s, let $\boldsymbol{\sigma}_q$ denote the per-channel standard deviations estimated from the training split. We perturb

$$\widetilde{q}_s = q_s + \boldsymbol{\eta} \odot \boldsymbol{\sigma}_q, \quad \eta_j \sim \mathcal{N}(0, \lambda^2). \tag{171}$$

Agent-source queries and map-source queries are perturbed separately for the corresponding interaction diagnostics. The experiment therefore measures sensitivity to normalized latent corruption and is interpreted separately from the paired simulator interventions.

F Additional Experimental Results and Analysis

This section provides extended empirical analyses of CIB. We first examine whether the observed gains persist under matched controls and independent training seeds. We then report representation probing and controlled-intervention results, followed by ablations of the physical parameterization and its constraint ranges. Finally, we analyze upstream task performance, computational overhead, iterative refinement stability, and geometric integrity.

F.1 Matched Controls and Statistical Robustness

The aggregate three-seed results are reported in Table 2 of the main paper. This appendix records the statistical protocol used for those comparisons. Each seed independently executes both training

Table 9: **Controlled appearance and agent interventions.** NIS measures response to appearance-only changes. Reaction Sensitivity measures the magnitude of the trajectory response to a physical intervention and is therefore reported without a monotonic direction.

Method	NIS (m) ↓				Agent Removal		Speed Intervention RS (m)	
	Cloud	Rain	Fog	Avg.	RS (m)	FBR (%) ↓	−2 m/s	+2 m/s
SparseDrive-P	0.10	0.23	0.31	0.21	0.16	27.1	0.13	0.17
+DSB	0.08	0.17	0.24	0.16	0.35	15.6	0.24	0.29
+CIB	**0.04**	**0.09**	**0.13**	**0.09**	**0.61**	**7.2**	**0.38**	**0.44**

Table 10: **Sensitivity to physical control bounds.** Excessively tight limits reduce expressivity, whereas overly relaxed limits weaken the physical constraint.

$a_{\max}$	$\kappa_{\max}$	L2 Avg. (m) ↓	CR Avg. (%) ↓	PDMS ↑
3	0.10	0.556	0.069	85.6
6	**0.20**	**0.546**	**0.064**	**86.4**
9	0.30	0.551	0.067	86.0
12	0.40	0.559	0.072	85.2
Unconstrained		0.571	0.080	83.9

stages, and all controlled variants initialized from a given seed share the corresponding Stage 1 checkpoint. Mean and standard deviation are computed across the three complete training runs. The training schedule is specified in Appendix C.6, and the distinction between training-seed variability and route-level bootstrap uncertainty is given in Appendix D.8.

F.2 Route-Level Uncertainty on Bench2Drive

Table 7 reports paired route-level bootstrap intervals for the three controlled Bench2Drive comparisons. The intervals are computed from the same 220-route evaluation outputs used in Table 1 and therefore require no additional model training.

F.3 Representation Probing and Controlled Interventions

Table 8 reports the probe protocol defined in Appendix E.2. Appearance decodability and physical-state decodability are interpreted jointly: lower appearance probe performance is informative only if task-relevant physical information remains recoverable. The mode-prior probe is reported separately because $p_s^{(k)}$ forms an explicit source-dependent pathway into the attention logits.

Table 9 reports the paired appearance and physical-agent interventions. NIS measures stability to rendering changes that preserve the physical scene, whereas Reaction Sensitivity measures the magnitude of the response when the traffic interaction itself changes. Neither metric is interpreted in isolation. The analysis focuses on selective sensitivity: reduced response to controlled appearance changes together with retained response to behaviorally relevant physical changes.

F.4 Ablation of Physical Parameterization

The core comparison among direct-state, yaw-rate, curvature, and class-aware mediators is reported in Table 2. The direct-state bottleneck already improves over the unconstrained host, while explicit physical integration provides a further gain. The curvature and class-aware variants perform similarly, suggesting that the main benefit is associated with physically structured mediation rather than with a uniquely optimal dynamic parameterization.

Because this comparison does not determine universally optimal physical bounds, Appendix F.5 separately evaluates sensitivity to the acceleration and curvature limits.

F.5 Sensitivity to Physical Bounds

We vary the acceleration and curvature limits while keeping the remaining training and evaluation settings fixed. The experiment tests the trade-off between two failure modes: overly restrictive bounds can exclude valid urban maneuvers, whereas very loose bounds reduce the effective strength

Table 11: **Upstream task performance on the nuScenes validation set.** The purpose of this analysis is to verify that physical mediation does not degrade the shared perception and prediction backbone.

Method	Detection		Tracking	Mapping	Motion
	NDS ↑	mAP ↑	AMOTA ↑	mAP ↑	minADE ↓
SparseDrive-P	0.522	0.411	0.381	0.556	0.624
PhysDrive-P	0.529	0.420	0.384	0.558	0.619

Table 12: **Computational cost of the matched-control variants.** All measurements use the same hardware, batch size, precision mode, and input resolution. DSB is included to verify that the matched-control result is not explained by a substantial capacity difference.

Method	Train Mem. (GB)	Infer Mem. (GB)	FLOPs (G)	Params (M)	FPS
SparseDrive-P	19.1	1.26	192.7	85.9	8.4
+DSB	30.6	1.32	290.8	98.4	7.3
PhysDrive-P	31.3	1.33	294.7	98.6	7.1

of the physical constraint. Table 10 reports the resulting sensitivity to the acceleration and curvature limits.

The default configuration is interpreted as an operating point within this trade-off rather than as a universal physical limit. These bounds specify the representation used by CIB and are not intended to reproduce the mechanical limits of every vehicle platform.

F.6 Upstream Task Preservation

Although CIB modifies cross-entity interaction downstream of sparse perception, its mediator losses backpropagate through shared network components. We therefore report detection, tracking, mapping, and motion prediction to check whether improved planning behavior is accompanied by a substantial degradation of upstream scene understanding. Table 11 reports the corresponding detection, tracking, mapping, and motion-prediction results.

These metrics serve only as a preservation check and are not interpreted as direct evidence about the information removed or retained by the interaction bottleneck.

F.7 Computational Overhead

Mode-expanded attention retains multiple physical hypotheses until the target query evaluates them and therefore introduces additional memory and computation relative to the host model. Table 12 reports training memory, inference memory, FLOPs, parameter count, and throughput under a common measurement setup.

We treat this increase as an explicit accuracy–efficiency trade-off. CIB operates on sparse interaction tokens and does not replicate the image backbone, but its overhead is not described as negligible.

F.8 Progressive Stability in Iterative Refinement

We evaluate refinement accuracy and inter-stage trajectory change jointly. Planning error measures whether later stages improve the trajectory, while Step Divergence measures how strongly the prediction changes relative to the initial refinement stage. Mean absolute jerk provides an additional trajectory-smoothness diagnostic. Table 13 reports planning error, Step Divergence, and mean absolute jerk jointly across refinement stages. The stagewise diagnostic predictions are auxiliary refinement outputs and do not coincide with the final trajectory reported in Table 1.

Low Step Divergence alone is not desirable if the model simply stops refining. We therefore interpret stability only when it is accompanied by competitive or improved planning accuracy.

Table 13: **Progressive refinement quality and stability.** Planning error and Step Divergence are reported jointly so that low divergence cannot be explained by the absence of useful refinement.

Method	L2 Avg. (m) ↓			SD at 3 s ↓		MAJ ↓
	Step 1	Step 3	Step 6	Step 3	Step 6	m/s^3
SparseDrive-I	0.62	0.64	0.69	0.45	0.88	4.25
PhysDrive-I	0.58	0.56	**0.55**	0.10	**0.24**	**1.75**

Table 14: **Curvature variation of predicted map geometry.** CVM is recomputed using the formal metric in Appendix D.7 and has units of m^{-2}. The adverse-appearance column averages the Cloudy, Rain, and Fog conditions defined in Table 6.

Method	Nominal ↓	Adverse appearance ↓
SparseDrive-P	0.031	0.064
PhysDrive-P	**0.027**	**0.039**
SparseDrive-I	0.034	0.073
PhysDrive-I	**0.026**	**0.038**

F.9 Geometric Smoothness and Latent Perturbation Robustness

We evaluate Curvature Variation per Meter using Appendix D.7. CVM measures local smoothness of the reconstructed polyline and is complementary to standard mapping accuracy; it is not a measure of semantic or topological correctness. We additionally retain normalized latent perturbation as a secondary stress test. This experiment asks whether corruption of source latent features is attenuated before affecting the planner. It complements, but does not replace, the paired CARLA interventions, which provide a cleaner test of controlled appearance sensitivity. Tables 14 and 15 report the geometric-smoothness and latent-corruption diagnostics, respectively.

G Cross-Architecture and Complementary Intervention Comparisons

This section evaluates CIB along two complementary dimensions. We first compare CIB with the Sparse Causal Intervention Scheme, denoted as SCIS, from CausalVAD (Tang et al., 2026) under matched SparseDrive backbones and study whether the two mechanisms can be combined. We then integrate CIB into the substantially stronger SparseDriveV2 architecture (Sun et al., 2026) to evaluate cross-architecture transfer.

The two studies address different questions. The SCIS comparison distinguishes explicit latent-context intervention from physically mediated communication. The SparseDriveV2 experiment evaluates whether CIB remains applicable beyond the SparseDrive variants used in our primary controlled study.

G.1 Relation to Sparse Causal Intervention

CausalVAD introduces the Sparse Causal Intervention Scheme, or SCIS, as a backdoor-inspired latent intervention method (Tang et al., 2026). SCIS constructs prototype dictionaries from learned object, map, and agent representations and uses them to adjust latent context during training.

CIB modifies a different part of the representation pipeline. It does not construct an explicit dictionary of contextual prototypes. Instead, it restricts source-to-target communication so that source-specific interaction content is generated from a physically grounded mediator representation.

The two approaches therefore involve different assumptions and costs. SCIS depends on an offline prototype representation of contextual variation, whereas CIB requires explicit physical-state supervision and additional interaction computation. We compare them under matched host architectures to study empirical behavior, not to rank their causal validity.

Table 15: **Robustness to normalized latent perturbations.** The experiment measures sensitivity to latent corruption and is not interpreted as a real-world causal intervention.

Method	λ	Ego–Agent		Ego–Map	
		L2 (m) ↓	CR (%) ↓	L2 (m) ↓	CR (%) ↓
SparseDrive-S	0.0	0.61	0.12	0.61	0.12
	0.5	0.66	0.22	0.75	0.13
	0.9	0.86	1.35	1.84	2.15
PhysDrive-S	0.0	0.56	0.08	0.56	0.08
	0.5	0.57	0.10	0.58	0.09
	0.9	0.62	0.23	0.78	0.31

Table 16: **Same-backbone comparison with SCIS on nuScenes.** Results marked with † are reported by the CausalVAD supplementary material.

Paradigm	Method	L2 Avg. (m) ↓	CR Avg. (%) ↓
Parallel	SparseDrive-P	0.58	0.08
	+ SCIS†	0.53	0.06
	+ CIB	0.54	0.06
	+ SCIS + CIB	**0.52**	**0.05**
Iterative	SparseDrive-I	0.62	0.10
	+ SCIS†	0.55	0.07
	+ CIB	0.56	0.07
	+ SCIS + CIB	**0.53**	**0.05**

G.2 Same-Backbone Comparison and Composition

We compare SCIS and CIB using the parallel and iterative SparseDrive configurations, for which the CausalVAD supplementary material reports SCIS results on the same host families. Reported literature values are marked separately from results reproduced under our implementation. Table 16 summarizes the same-backbone comparison and the composition of SCIS with CIB.

For the combined model, SCIS is applied only to the latent-query component of the source descriptor:

$$\widetilde{q}_s = \mathrm{SCIS}\left(q_s; \mathcal{Z}\right), \tag{172}$$

$$\widetilde{\chi}_s = \left(\widetilde{q}_s, b_s, \tau_s\right), \quad \mathcal{M}_s = \mathcal{G}\left(\widetilde{\chi}_s\right), \tag{173}$$

where $\mathcal{Z}$ denotes the SCIS prototype dictionary. The subsequent CIB interaction uses only the resulting mediator representation as its source-dependent input.

This composition preserves the source-side no-bypass property of the CIB edge provided that SCIS output is not introduced through any additional key, value, or attention-bias path. If the combined model improves over either component alone, we interpret the result as evidence of empirical complementarity under the tested architecture. If it does not, we make no stronger claim about whether the two methods remove the same information.

G.3 Comparison under Controlled Interventions

Standard planning metrics do not reveal whether SCIS and CIB alter model sensitivity in the same way. We therefore evaluate the methods under the paired appearance and physical-agent interventions defined in Appendix E. Table 17 compares the corresponding sensitivity profiles under controlled appearance and agent interventions.

Appearance-only pairs measure stability to controlled rendering changes, whereas agent interventions modify the traffic interaction state. We report the corresponding metrics side by side rather than collapsing them into a single aggregate robustness score. Differences between the methods are interpreted as empirical sensitivity profiles under the tested interventions, not as estimates of causal effects.

G.4 Integration with SparseDriveV2

The aggregate Bench2Drive results are reported in Table 18. Here we describe the architectural substitution in detail.

Table 17: **Mechanism comparison under controlled interventions.** NIS measures sensitivity to appearance-only changes. Reaction Sensitivity measures the trajectory response to removing a behaviorally relevant agent and is reported without a monotonic direction.

Method	NIS Avg. (m) ↓	Removal RS (m)	FBR (%) ↓
SparseDrive-P	0.21	0.16	27.1
+ SCIS	0.13	0.32	13.9
+ CIB	0.09	0.61	7.2
+ SCIS + CIB	**0.07**	**0.65**	**5.8**

Table 18: **Cross-backbone transfer on Bench2Drive.** The reported SparseDriveV2 result is quoted from the original work; reproduced rows are evaluated with our Bench2Drive v0.0.4 setup.

Method	Source	DS ↑	SR (%) ↑
SparseDriveV2[†]	Reported	89.15	70.00
SparseDriveV2	Reproduced	89.0	69.5
SparseDriveV2 + CIB	Reproduced	**89.9**	**71.5**

SparseDriveV2 factorizes planning trajectories into path and velocity components and applies coarse-to-fine candidate scoring. We preserve this trajectory representation and replace only the agent/map-to-planning interaction interfaces with CIB-mediated communication. Agent and map queries are converted to their corresponding mediator representations before contributing source-dependent content to the path and velocity embeddings. Candidate filtering, fine-grained scoring, trajectory re-conditioning, and trajectory-to-control conversion remain unchanged.

Both the reproduced SparseDriveV2 baseline and SparseDriveV2+CIB use the same Bench2Drive v0.0.4 training data, sensor configuration, route split, and evaluation seeds. The experiment therefore measures the effect of the interaction substitution rather than a change in the planning vocabulary or evaluation protocol.

G.5 Scope of the Cross-Method Comparison

The experiments in this section address mechanism complementarity and architectural transfer. SCIS modifies latent context through learned prototypes, whereas CIB constrains cross-entity communication through physically grounded mediators. Their combination tests whether these two design choices provide complementary empirical benefits under a common host architecture.

We do not use these comparisons to define a hierarchy of causal validity. Likewise, improved performance after transferring CIB to SparseDriveV2 supports architectural portability but does not strengthen the front-door interpretation. Throughout the paper, we keep causal motivation, architecture-level guarantees, and empirical robustness as distinct levels of evidence.