Intervene3D: Intervention-Based Controlled Inference for Multimodal Perception under Partial Observability

Constantino Msigwa[1] **Denis Bernard**[1] **Jaeseok Yun**[1,2*]

[1]Department of Future Convergence Technology, Soonchunhyang University,
22 Soonchunhyang-ro, Asan, 31538, South Korea
[2]Department of Internet of Things, Soonchunhyang University,
22 Soonchunhyang-ro, Asan, 31538, South Korea

Abstract

Multimodal 3D detection becomes unreliable under partial observability caused by occlusion, LiDAR sparsity, and camera–LiDAR miscalibration. Most fusion detectors are fully feed-forward: once trained, they provide no explicit mechanism to incorporate structured test-time constraints or to bound how such constraints can alter predictions. We propose **Intervene3D**, a controlled-inference wrapper for BEV-based detectors that performs *metric-bounded latent refinement* at inference time. Given sensor inputs, the detector produces an initial BEV latent $\mathbf{z}_0$ and an evidence-conditioned *trust metric* $\mathbf{H}$ (a diagonal precision derived from predicted uncertainty). A separate control stream (predicted from sensor inputs in our main experiments, with external control demonstrated only preliminarily) is mapped to a fully specified differentiable constraint energy (count intervals and spatial admissibility). Inference then updates $\mathbf{z}$ using $K \in \{1, 2, 3\}$ *diagonal-metric preconditioned* projected steps that reduce constraint energy while remaining within an uncertainty-gated trust region defined by $\mathbf{H}$. To improve robustness to extrinsic drift, we additionally introduce **phase-aware spectral alignment** that normalizes cross-modal Fourier-phase discrepancies prior to attention. We report clean accuracy and a severity-swept robustness protocol (robustness-AUC, feasibility diagnostics, and conflict-induced FP inflation).

1 Introduction

Multimodal 3D detection is routinely deployed where sensor evidence is incomplete or unreliable, e.g., objects occluded in RGB, sparse returns in LiDAR, or mild camera–LiDAR miscalibration. In such settings multiple latent scene explanations can match the same measurements, and a purely feed-forward fusion network must implicitly choose one explanation using priors encoded in its parameters [21, 23]. A key limitation of current multimodal detectors is that inference is *fully amortized*: predictions are produced by a single forward pass $Y = F_\theta(X)$ [33, 24, 14, 47, 42]. When external information is available at test time (e.g., map priors, expected object density, restricted regions, or reliability signals), existing approaches typically inject it by conditioning the encoder, e.g., concatenation or FiLM, yielding $Z = f_\theta(X, T)$. This entangles control with feature construction and offers no explicit mechanism to (i) enforce structured constraints, (ii) attenuate control when evidence is strong, or (iii) audit how much control changes predictions. We address this missing abstraction by introducing an explicit *inference-time refinement operator* for BEV detectors. Intervene3D separates (a) **evidence**, which defines an initial BEV latent $\mathbf{z}_0$ and an evidence-conditioned trust metric, from (b) **control**, which defines differentiable constraint energies. Inference performs a small number of

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

bounded updates to $\mathbf{z}$ that reduce constraint violations while remaining within an uncertainty-gated trust region. This design makes test-time intervention measurable (sensitivity vs. uncertainty) and enables stress tests under conflicting control.

In addition, we improve robustness to mild calibration drift using phase-aware spectral alignment in the frequency domain, motivated by the Fourier shift theorem. We evaluate Intervene3D on KITTI and nuScenes using clean metrics and severity-swept robustness curves, including robustness-AUC and conflict-induced false-positive inflation. The primary contributions of our method are as follows:

- **Controlled inference interface for BEV detectors.** We introduce an inference-time operator that refines BEV latents under an evidence-conditioned trust region, enabling structured test-time intervention without re-encoding sensor evidence.
- **Explicit, reproducible constraints.** We define differentiable constraint energies for (i) cardinality intervals and (ii) spatial admissibility fields directly on BEV logits for stable gradients.
- **Phase-aware spectral alignment.** We propose a frequency-domain alignment module that estimates and compensates for cross-modal phase discrepancies prior to attention to improve resilience under extrinsic perturbations.
- **Auditable diagnostics and robustness protocol.** We specify severity-swept corruption benchmarks (robustness-AUC), feasibility diagnostics for the trust-region constraint, and conflict-induced FP inflation.

2 Related Work

Multimodal BEV perception. LiDAR-camera 3D detectors commonly use early, mid-level, or BEV-centric fusion to combine geometric and semantic cues [24, 40, 47, 48]. These methods achieve strong performance by mapping observations to a BEV latent $Z = f_\theta(X)$ and predicting detections with $Y = h(Z)$ [10, 15, 9, 11]. However, inference remains fully amortized: sensor evidence, uncertainty, and learned priors are absorbed into one feed-forward pass. Thus, they lack an explicit mechanism for test-time constraint satisfaction or posterior-like refinement when observations are sparse, misaligned, or contradictory.

Frequency-domain robustness and completion. Frequency-aware representations improve robustness by separating global low-frequency structure from high-frequency details, but they are usually applied within each modality and reintegrated through standard fusion [4, 26, 45, 13, 30, 25]. Similarly, generative completion methods reconstruct dense BEV, occupancy, or image features from sparse observations [29]. Although useful for plausibility, these methods optimize reconstruction-style objectives and remain feed-forward, without a bounded mechanism for task-driven latent intervention.

Prompting and conditional adaptation. Prompt tokens, FiLM layers, and other conditioning mechanisms inject task or context information into visual models, often written as $Z = f_\theta(X, T)$ [37, 5, 46, 16, 31]. While this improves flexibility, control signals are entangled with sensor evidence during encoding. Consequently, their influence is difficult to bound, attenuate, or audit, especially when prompts conflict with reliable observations.

Test-Time 3D Detection Correction. Zhang et al. [46] explore visual prompting for online 3D detection. In contrast, Intervene3D uses uncertainty-adaptive, trust-region-constrained latent refinement with evidence-conditioned bounds and projected optimization of differentiable constraint energies. This bounds latent interventions but does not guarantee detection correctness or downstream safety. A direct comparison under our protocol is in Supplement Sec. N.3.

2.1 Missing Abstraction: Metric-Bounded Controlled Inference

Existing fusion, completion, and conditioning methods mostly perform single-pass amortized inference. When control variables are used, they are injected into the encoder, making it difficult to enforce test-time constraints, preserve reliable evidence, or measure how much control changed the prediction. We instead separate evidence and control: the evidence-only encoder produces $z_0 = \phi(x)$ and an uncertainty-derived metric $H(x)$, while structured control τ defines a differentiable constraint energy $\mathcal{C}(z; \tau)$. Controlled inference solves $z^*(x, \tau) = \arg\min_z \ \mathcal{C}(z; \tau) \quad \text{s.t.} \ (z - z_0)^\top H(x)(z - z_0) \leq$

$2\delta(u)$. where $\delta(u)$ is an uncertainty-adaptive intervention budget. This yields an auditable interface: evidence fixes the initial latent and trust metric, while control can only refine the latent within a bounded region. This trust region can be interpreted as a local KL bound under the Gaussian surrogate $p(z|x) = \mathcal{N}(z_0, H^{-1})$ with fixed covariance. We do not assume a true generative posterior; the surrogate is used only to bound test-time intervention. Intervene3D therefore contributes not a new optimizer, but a controlled inference interface for BEV perception with bounded influence, feasibility diagnostics, and conflict-aware false-positive analysis.

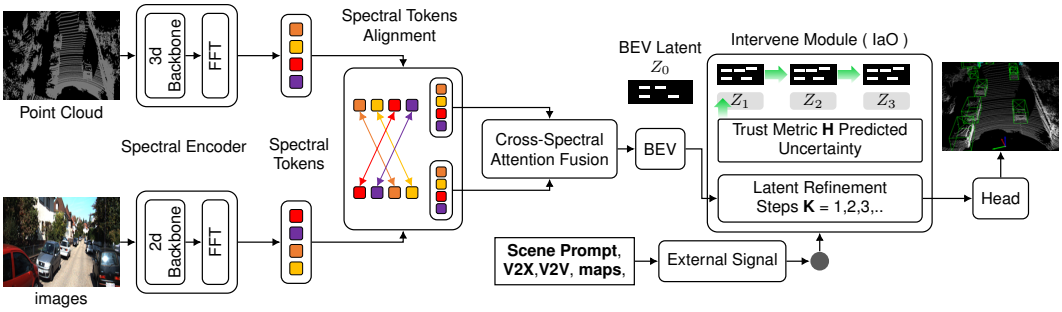


Figure 1: **Intervene3D.** Camera and LiDAR inputs produce an evidence latent $\mathbf{z}_0$ and uncertainty metric $\mathbf{H}$. Structured control defines a constraint energy, and Intervene3D refines the latent within an uncertainty-gated trust region before final 3D detection.

3 Method

Overview of the Approach. Intervene3D formulates multimodal 3D detection under partial observability as *controlled inference*. We separate *evidence* (sensor-derived context) from *control* (structured constraints), enabling test-time intervention without modifying sensor features. RGB and LiDAR features are encoded in the spectral domain and aligned via cross-spectral attention to form an evidence-only latent context. Separately, structured prompts are encoded as a control stream and applied through an intervention operator that refines BEV latents in a constrained manner. This decoupling lets control guide inference under ambiguity while preserving attribution to observed evidence.

Conditioning vs. inference. Standard prompt or FiLM-based conditioning implements a larger feed-forward mapping $Z = f_\theta(X, T)$, entangling control with evidence during encoding. Such architectures lack an explicit mechanism to enforce constraint satisfaction or to bound the influence of T based on evidence reliability. In contrast, Intervene3D introduces an inference-time operator (IaO) that performs structured latent refinement under explicit constraints. Control therefore modifies the inference process rather than the feature construction pathway, yielding an intervenable and auditable interface for test-time posterior adjustment.

3.1 Spectral Encoders

To improve robustness to spatial misalignment and calibration noise, we encode both RGB and LiDAR features in the frequency domain. Our design follows the same principle for each modality: (i) extract spatial/voxel features with a task-appropriate backbone, (ii) transform features into spectral space (FFT/DCT), (iii) summarize band-limited components as *spectral tokens*, and (iv) predict an uncertainty map to quantify sensor reliability under occlusion and sparsity.

Image spectral encoder. Given an RGB image $\mathbf{I} \in \mathbb{R}^{3\times h\times w}$, we extract intermediate feature maps using a configurable 2D CNN backbone : $\mathbf{X}_I = f_I(\mathbf{I}) \in \mathbb{R}^{C_I\times H'\times W'}$. To obtain a stable feature dimension across backbones, we project to C channels using a 1×1 convolution followed by GroupNorm: $\hat{\mathbf{X}}_I = \mathrm{Proj}_I(\mathbf{X}_I) \in \mathbb{R}^{C\times H'\times W'}$. We then convert features to the frequency domain via a 2D Fourier transform with frequency centering: $\mathcal{F}_I = \mathrm{fftshift}\Big(\mathrm{FFT2}(\hat{\mathbf{X}}_I)\Big)$, and represent the spectral signal by concatenating magnitude and phase, $\mathbf{F}_I = \big[\ |\mathcal{F}_I|\ ;\ \angle\mathcal{F}_I\ \big] \in \mathbb{R}^{2C\times H'\times W'}$. Finally, we partition the centered spectrum into B concentric radial frequency bands.

Let $\mathbf{M}_b \in \{0,1\}^{H' \times W'}$ denote the binary mask for band b, defined by an annulus with inner radius r_b and outer radius r_{b+1}. We compute a band token by masking, projecting, and global pooling: $\mathbf{t}_I^{(b)} = \mathrm{GAP}\Big(\mathrm{Conv}_b\big(\mathbf{F}_I \odot \mathbf{M}_b\big)\Big) \in \mathbb{R}^d, \mathbf{T}_I \quad = \big[\mathbf{t}_I^{(1)}, \ldots, \mathbf{t}_I^{(B)}\big] \in \mathbb{R}^{B \times d}$. which matches our implementation: circular masks (radial bins), per-band 1×1 projections, and adaptive average pooling.

LiDAR spectral encoder. For LiDAR, we start from voxel features produced by a voxel feature encoder (VFE): $\mathbf{X}_L \in \mathbb{R}^{N \times C_L}, \mathbf{V} \in \mathbb{Z}^{N \times 4}$, where N is the number of non-empty voxels and $\mathbf{V}$ stores voxel coordinates (batch index and 3D grid indices). To obtain a frequency-domain representation, we reconstruct a coarse 3D grid $\mathbf{G}_L \in \mathbb{R}^{C_L \times S \times S \times S}$ (with modest S for efficiency) by scattering voxel features into their grid locations, and apply a 3D Fourier transform: $\mathcal{F}_L = \mathrm{FFTN}(\mathbf{G}_L), \qquad \mathbf{F}_L = \big[\, |\mathcal{F}_L| \,;\, \angle\mathcal{F}_L \,\big]$. We then extract B band-wise LiDAR tokens using lightweight linear projections (one per band) applied to the globally pooled spectral features inside the radial shell $\mathbf{M}_b^L$ of the 3D spectrum: $\mathbf{t}_L^{(b)} = \mathbf{W}_L^{(b)}\, \mathrm{GAP}\big(\mathbf{F}_L \odot \mathbf{M}_b^L\big), \mathbf{T}_L = \big[\mathbf{t}_L^{(1)}, \ldots, \mathbf{t}_L^{(B)}\big]$, where each $\mathbf{W}_L^{(b)}$ reduces dimensionality to form compact per-band descriptors. As an alternative (when 3D FFT is undesirable), we support a learnable Fourier projection that maps voxel features into sinusoidal components before band projection .

Uncertainty estimation. Both encoders additionally predict uncertainty to characterize modality reliability. For spatial image features, we use a shallow conv head: $\mathbf{U}_I = \sigma\Big(h_I(\hat{\mathbf{X}}_I)\Big) \in [0,1]^{C \times H' \times W'}$, and for voxel-wise LiDAR features we use an MLP head: $\mathbf{U}_L = \sigma(h_L(\mathbf{X}_L)) \in [0,1]^{N \times C_L}$, implemented by `UncertaintyHead`, which switches between conv (4D tensors) and linear (2D tensors). These uncertainty signals are later pooled and used to modulate cross-modal fusion and BEV synthesis.

Phase-Aware Spectral Alignment. Standard spectral fusion applies attention directly to magnitude–phase tokens. However, spatial misalignment between LiDAR and camera sensors induces phase shifts in the Fourier domain. According to the Fourier shift theorem, a spatial translation δ corresponds to a multiplicative phase term $e^{-j\omega\delta}$. If uncorrected, such phase discrepancies degrade cross-modal alignment. To address this, we introduce phase-aware spectral alignment that explicitly normalizes relative phase offsets prior to cross-modal interaction.
Complex spectral representation. Let the spectral representations of image and LiDAR features be expressed in polar form: $\mathcal{F}_I = A_I e^{j\phi_I}, \mathcal{F}_L = A_L e^{j\phi_L}$, where A denotes magnitude and ϕ denotes phase.

Relative phase estimation. We estimate a band-wise relative phase offset $\Delta\phi$ using a lightweight predictor operating on concatenated magnitude statistics: $\Delta\phi = f_\Delta(\mathrm{GAP}(A_I), \mathrm{GAP}(A_L))$. This predictor models systematic phase shifts caused by calibration noise or small spatial perturbations; we treat it as a learned feature-alignment mechanism in a shared feature representation, not as an exact extrinsic calibration. **Isolating phase correction.** We compare spectral tokens without phase correction ($\Delta\phi = 0$), phase correction only (applied before spatial-domain fusion), and full phase-aware spectral attention under controlled extrinsic perturbations (Sec. 5).

Band tokenization. Corrected spectral tensors are partitioned into B radial frequency bands and converted into tokens: $\mathbf{T}_I = \mathrm{Tokenize}(A_I, \phi_I), \mathbf{T}_L = \mathrm{Tokenize}(A_L, \phi_L - \Delta\phi)$. Tokens are projected into a shared embedding space.
Cross-modal interaction. We then apply multi-head cross-attention between the aligned spectral tokens: $\mathbf{T}_{\text{fusion}} = \mathrm{MHA}(\mathbf{T}_I, \mathbf{T}_L, \mathbf{T}_L) + \mathrm{MHA}(\mathbf{T}_L, \mathbf{T}_I, \mathbf{T}_I)$. **Uncertainty-aware weighting.** Global uncertainty descriptors derived from spatial reliability maps are used to compute a fusion coefficient $\omega \in (0,1)$. The fused tokens are scaled as: $\mathbf{T}_{\text{fusion}} \leftarrow \omega \cdot \mathbf{T}_{\text{fusion}}$. This prevents unreliable modalities from dominating alignment.

3.2 Source of Semantic Control and Constraint Generation

Intervene3D treats semantic control as an intervention signal that modifies the *inference trajectory* while leaving sensor encoding unchanged. Control is represented by an abstract variable $\tau \in \mathcal{T}$, which may be exogenous (maps, rules, V2X) or internally predicted when exogenous priors are unavailable. A learned constraint generator maps τ into a differentiable constraint specification:

$\mathcal{G}_\psi : (\tau, c) \to \mathcal{S}$, where c denotes contextual BEV features (detached from gradients w.r.t. τ-prediction when applicable), and $\mathcal{S}$ parameterizes the constraint energy $\mathcal{C}(z; \mathcal{S})$ used by IaO (Sec. 3.3).

Exogenous control. When external priors are available, τ is provided by modules causally upstream or external to perception (e.g., HD maps, route planners, policy engines, weather metadata). In this case, τ is treated as exogenous and can be modified at inference time (an intervention $do(\tau = t)$) without retraining.

Internally predicted control. In the main experiments, external priors (e.g., HD maps, V2X, route information, or manual descriptions) are unavailable; thus, we instantiate τ as a predicted scene-description prompt derived solely from sensor inputs at validation and test time. To avoid circular dependency between sensor noise and control, the control predictor Φ adopts a spatially invariant global pooling architecture. Unlike the 3D detection backbone, which depends on precise local geometric alignment and is sensitive to drift and occlusion, Φ captures global scene statistics (e.g., object counts), enabling robust control estimation.

Control Signal Construction and Supervision. The control signal τ is generated from RGB-LiDAR observations by a lightweight predictor Φ, without external information. It comprises class-presence probabilities, coarse count intervals $[m_c, M_c]$, spatial admissibility fields A_c, scene-layout descriptors, and reliability cues, which parameterize the constraint energy $\mathcal{C}(z; \mathcal{S})$. The predictor is pretrained using coarse supervision derived exclusively from training annotations and frozen during detector optimization. At inference, τ is predicted solely from sensor inputs, without access to validation or test annotations. Ground-truth control is used only in separate oracle upper-bound experiments.

Leakage risk and required non-leakage diagnostics. Internally predicted control can inadvertently encode dataset priors. Therefore, whenever using predicted τ, the following diagnostics should be reported. Table 6 covers the neutral (dropout), random, swap, and oracle checks; the cross-dataset check is not evaluated in this paper. **Dropout:** disable τ at test time ($\tau = \bar{\tau}$) and report clean/robustness deltas. **Swap:** swap τ across scenes; performance should degrade and should not improve. **Random:** draw τ from its marginal distribution; performance should not improve. **Cross-dataset Φ:** pretrain Φ on a disjoint split/dataset and test transfer. **Oracle upper bound:** use GT coarse summaries (same schema) to quantify headroom and separate leakage vs. mechanism.

Preventing Circular Information Flow. To prevent cyclic dependencies between control generation and latent refinement, we explicitly decouple the prompt predictor from the detector optimization. The predictor Φ is pretrained and frozen, and its output $\tau = \Phi(X)$ is detached before being passed to IaO.Consequently, gradients from the constraint energy do not update the control predictor. The evidence encoder is executed only once to obtain the initial latent representation z_0, and the structured control signal remains fixed throughout all refinement iterations. In particular, τ is never regenerated from the refined latent z_k. This design prevents iterative feedback between control prediction and latent refinement while preserving a fixed sensor-derived reference for the intervention.

Constraint energies (fully specified). Let the detection head produce BEV class logits $L_c(z) \in \mathbb{R}^{H_b \times W_b}$. Define soft BEV occupancy $P_c(z) = \sigma(L_c(z)) \in (0,1)^{H_b \times W_b}$. The constraint generator outputs $\mathcal{S} = \left\{ [m_c, M_c],\ A_c,\ \alpha_c, \beta_c \right\}_{c=1}^{C}$, where $[m_c, M_c]$ is a *soft-count interval*, $A_c \in [0,1]^{H_b \times W_b}$ is a *spatial admissibility field*, and $\alpha_c, \beta_c \geq 0$ are weights

(i) Count-interval loss. Define soft mass $N_c(z) \triangleq \sum_{i=1}^{H_b} \sum_{j=1}^{W_b} P_c(z)_{ij}$. Then

$$\mathcal{L}_c^{\text{count}}(z) = \big[\text{ReLU}(m_c - N_c(z))\big]^2 + \big[\text{ReLU}(N_c(z) - M_c)\big]^2. \tag{1}$$

To prevent latent drift, where optimization satisfies the constraints by degrading the latent representation, refinement is restricted to the metric-bounded trust region defined by H (Sec. 3.3). Because $\nabla_z \mathcal{C}(z)$ is computed through the soft occupancy $P_c(z)$, the update acts mainly along low-confidence (small-H) latent directions and moves high-precision directions only slightly. **(ii) Spatial admissibility loss.**

$$\mathcal{L}_c^{\text{spatial}}(z) = \frac{1}{H_b W_b} \sum_{i,j} \big(1 - A_c(i,j)\big)\, P_c(z)_{ij}. \tag{2}$$

(iii) Uncertainty-gated control weight. Let $u \in [0,1]$ be the global uncertainty scalar (defined in Sec. 3.3). We gate control influence by $w(u) = \text{clip}(u^\kappa,\ 0,\ 1)$, with $\kappa \geq 1$ (default $\kappa = 3$).

Total constraint energy. $\mathcal{C}(z; \mathcal{S}) = w(u) \sum_{c=1}^{C} \Big(\alpha_c\, \mathcal{L}_c^{\text{count}}(z) + \beta_c\, \mathcal{L}_c^{\text{spatial}}(z) \Big)$. **Conflict construction (for hallucination audits).** To generate prompt-evidence conflict, we perturb $[m_c, M_c]$ to

contradict observed soft masses (e.g., enforce $m_c > N_c(z_0) + \Delta$) and/or set A_c to suppress regions with high evidence support.

3.3 Intervention-as-Metric-Bounded Constrained Inference (IaO)

Intervene3D implements semantic control as an explicit *metric-bounded latent refinement* operator applied at inference time. Unlike FiLM/prompt concatenation, which implements a larger feed-forward mapping $f_\theta(x, \tau)$ and entangles control with evidence, IaO refines an evidence-derived latent z *post hoc* under differentiable constraints while explicitly bounding how far refinement may deviate from the evidence estimate.

Evidence trust metric. Given sensor inputs x, the backbone produces an initial BEV latent $z_0 = \phi(x)$. We predict diagonal log-variance $s(x) \in \mathbb{R}^D$ and define $H(x) = \mathrm{diag}\big(\exp(-s(x))\big) \succ 0$.

Trust-region refinement. Given $\mathcal{S}$, we refine by $\min_z \mathcal{C}(z; \mathcal{S}) \quad \text{s.t.} \quad \|z - z_0\|_H^2 \le 2\delta(u), \|v\|_H^2 \triangleq v^\top H v$, with $\delta(u) = \delta_{\min} + (\delta_{\max} - \delta_{\min})u^\gamma$. In the implementation $\|v\|_H^2$ is divided by the number D of latent elements, i.e., the budget is per element; all statements below hold unchanged with $\delta(u)$ replaced by $D\delta(u)$ for the unnormalized norm.

Scope of the Feasibility Guarantee. At each refinement iteration, IaO projects the updated latent representation onto the uncertainty-adaptive trust region $\mathcal{E}(x) = \left\{ z : \|z - z_0\|_{H(x)}^2 \le 2\delta(u) \right\}$. Consequently, the projected latent satisfies $\|z_k - z_0\|_{H(x)}^2 \le 2\delta(u), \forall k$, up to floating-point precision. This guarantee applies exclusively to latent-space intervention magnitude, with the same sample-specific trust metric $H(x)$ and intervention budget $\delta(u)$ used during projection. It does not guarantee that the decoded detections satisfy all semantic constraints, that predicted bounding boxes are correct, or that false positives cannot occur after nonlinear decoding and non-maximum suppression. We therefore evaluate latent feasibility separately from detection-level robustness and prompt-evidence conflict.

3.4 Uncertainty Budget and Projected Refinement

Global uncertainty scalar u (explicit). Let $\bar{U}_I$ be the spatial mean of image uncertainty and $\bar{U}_L$ the mean of LiDAR uncertainty: $\bar{U}_I = \frac{1}{CH'W'} \sum_{c,i,j} U_I(c,i,j), \bar{U}_L = \frac{1}{NC_L} \sum_{n,c} U_L(n,c)$, and define $u = \mathrm{clip}\big(\lambda_I \bar{U}_I + \lambda_L \bar{U}_L,\ 0,\ 1\big), \lambda_I + \lambda_L = 1$. We use $\lambda_I = \lambda_L = 0.5$ unless otherwise stated.

Local Gaussian surrogate (trust-region view). We do not assume a generative detector. Instead, we use $H(x)$ to define a *local quadratic surrogate* around z_0:$p(z \mid x) \triangleq \mathcal{N}(z_0, H^{-1})$. This surrogate provides a computable geometry (an H-norm) for bounding refinement.

Constraint-regularized refinement with an uncertainty-adaptive budget. Given the constraint specification $\mathcal{S}$ produced from control τ (Sec. 3.2) and the constraint energy $\mathcal{C}(z; \mathcal{S})$, IaO solves the trust-region problem subject to the constraint $\|z - z_0\|_H^2 \le 2\delta(u)$. Thus, interventions are larger only when evidence is less reliable.

Projected natural-gradient refinement (feasible for multi-step $K > 1$). We perform K refinement steps ($K \in \{1, 2, 3\}$ by default). Let $g_k = \nabla_z \mathcal{C}(z_k; \mathcal{S})$. We use the preconditioned direction (a natural-gradient step under the diagonal metric) $d_k = H^{-1} g_k$. We first take an unconstrained step $\tilde{z}_{k+1} = z_k - \eta_k d_k$ with $\eta_k \le \eta_{\max}$, then project back to the trust region:

$$\tilde{z}_{k+1} = z_k - \eta_k d_k, \quad z_{k+1} = \Pi_{\mathcal{E}}(\tilde{z}_{k+1}), \quad \mathcal{E} \triangleq \{ z : \|z - z_0\|_H^2 \le 2\delta(u) \}. \tag{3}$$

Closed-form metric projection. Let $\Pi_{\mathcal{E}}$ denote the projection onto $\mathcal{E}$ measured in the H-norm, i.e., $\Pi_{\mathcal{E}}(\tilde{z}) = \arg\min_{z \in \mathcal{E}} \|z - \tilde{z}\|_H$ (this is not the Euclidean projection onto the ellipsoid, which has no closed form in general). Because $\mathcal{E}$ is an H-ball centered at z_0, this projection is a radial rescaling. Let $v = \tilde{z}_{k+1} - z_0$. If $\|v\|_H^2 \le 2\delta(u)$, then $\Pi_{\mathcal{E}}(\tilde{z}_{k+1}) = \tilde{z}_{k+1}$. Otherwise, $\Pi_{\mathcal{E}}(\tilde{z}_{k+1}) = z_0 + \sqrt{\frac{2\delta(u)}{v^\top H v}}\, v$. This projection is inexpensive and guarantees feasibility.

Guarantee (Feasibility for all k). For any $k \ge 0$, the update in Eqs. (3) satisfies $\|z_{k+1} - z_0\|_H^2 \le 2\delta(u)$. *Proof.* By definition, $z_{k+1} = \Pi_{\mathcal{E}}(\tilde{z}_{k+1}) \in \mathcal{E}$, hence it satisfies the constraint.

KL interpretation (restricted family; optional). If we restrict to Gaussians with shared covariance $q(z) = \mathcal{N}(\mu, H^{-1})$ and $p(z \mid x) = \mathcal{N}(z_0, H^{-1})$, then $\mathrm{KL}(q\|p) = \frac{1}{2}(\mu - z_0)^\top H (\mu - z_0) = \frac{1}{2}\|\mu - z_0\|_H^2$. Thus, Eq. (3.3) can be interpreted as bounding $\mathrm{KL}(q\|p)$ *within this restricted Gaussian*

family. We use this only as an information-geometric interpretation of the metric-bound; we do not require a generative detector.

Decoding. After K steps, we decode $z^* \triangleq z_K$ into BEV features and apply the standard detection head: $M_{\mathrm{BEV}} = f_{\mathrm{dec}}(z^*), \hat{y} = h(M_{\mathrm{BEV}})$.

3.5 Training Objectives

Detection head. The refined latent z^* is decoded into BEV features $M_{\mathrm{BEV}} = f_{\mathrm{dec}}(z^*)$ and passed to the same 3D detection head as the baseline. This keeps the detector head fixed and isolates the effect of spectral alignment and IaO.

Objective. Intervene3D is trained with the base detection loss and lightweight auxiliary terms: $\mathcal{L} = \lambda_{\mathrm{det}}\mathcal{L}_{\mathrm{det}} + \lambda_{\mathrm{spec}}\mathcal{L}_{\mathrm{spec}} + \lambda_{\mathrm{prompt}}\mathcal{L}_{\mathrm{prompt}} + \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}}$. $\mathcal{L}_{\mathrm{det}}$ is unchanged from the base detector and includes standard classification and 3D box-regression losses.

Spectral consistency. To preserve spatial feature content after spectral alignment, we use $\mathcal{L}_{\mathrm{spec}} = \|\tilde{X}_I - \mathrm{sg}(\hat{X}_I)\|_1 + \|\tilde{X}_L - \mathrm{sg}(\hat{X}_L)\|_1$, where $\hat{X}_I, \hat{X}_L$ are spatial image/LiDAR features before spectral transform, $\tilde{X}_I, \tilde{X}_L$ are inverse-transformed reconstructions, and $\mathrm{sg}(\cdot)$ denotes stop-gradient. Thus, this loss trains the spectral modules without distorting the task backbone.

Prompt stability. For a predicted scene-description prompt τ and a perturbed prompt $\tau' = \tau + \Delta\tau$, we regularize IaO sensitivity by $\mathcal{L}_{\mathrm{prompt}} = \|z^*(x,\tau) - \mathrm{sg}(z^*(x,\tau'))\|_H^2$. This reduces brittle dependence on small prompt errors.

Refinement consistency. For a neutral prompt $\bar{\tau}$ with weak constraints, we use $\mathcal{L}_{\mathrm{rec}} = \|z^*(x,\bar{\tau}) - \mathrm{sg}(z_0)\|_H^2$, which prevents unnecessary drift from the evidence latent z_0.

Implementation. $H(x)$ is computed once per sample and reused in all H-norm terms. For predicted control, the prompt predictor Φ is frozen and its output is detached. Validation/test annotations are never used to construct τ; oracle prompts are used only for upper-bound diagnostics. Auxiliary terms are ablated in Sec. 5.

Training and Inference Separation. The structured-control predictor and detector are optimized in separate stages. First, Φ is pretrained using annotation-derived coarse targets from the training split. During subsequent detector training, Φ remains frozen and its predictions are detached before IaO refinement. The detector is optimized using the detection objective and the auxiliary spectral, prompt-stability, and refinement-consistency losses defined above. At inference time, sensor evidence is encoded once, the predicted control is generated once, and IaO performs K bounded refinement steps without updating model parameters or regenerating the control signal. This separation ensures that the control branch does not receive gradients from the detector's refinement objective.

Table 1: **3D Object Detection Results on the KITTI Validation Set using R40.** 3D mAP (%) for vehicles, pedestrians, and cyclists under Easy, Moderate, and Hard difficulty levels.

Method	Mod	Year	Vehicle mAP			Pedestrian mAP			Cyclist mAP		
			Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
CSMNET[27]	LC	2024	89.02	81.95	77.63	48.95	42.15	39.44	82.86	67.65	60.90
GFA-SMT [17]	LC	2024	89.12	80.24	74.52	50.36	43.18	41.63	79.82	65.44	60.21
SFD-IAFNet[12]	LC	2025	**95.84**	88.90	87.67	-	-	-	-	-	-
TransLGS [28]	LC	2025	88.47	79.35	73.86	49.15	40.58	39.68	78.82	65.75	59.48
FFAFormer[8]	LC	2026	92.23	87.44	81.21	-	-	-	-	-	-
3DifFusionDet[35]	LC	2026	91.60	89.50	82.60	-	-	-	-	-	-
BEVFormer++ [20]	LC	2026	90.14	82.02	78.31	49.35	43.22	39.93	83.75	**68.56**	61.14
Intervene3D w/o IaO	LC		95.58	89.56	87.92	68.28	59.79	53.38	86.59	67.88	64.11
Intervene3D	LC		95.33	**89.75**	**88.00**	**69.37**	**62.59**	**53.73**	**87.95**	68.14	**65.78**

4 Experiments and Results

4.1 Experimental Setup

Datasets and metrics. On KITTI, we use the standard 3,712/3,769 train/val split and report AP_{3D} for Car under Easy, Moderate, and Hard settings, with additional multi-class results for Car, Pedestrian,

Table 2: 3D mAP(%) results for the Car on the KITTI `test` set using R40.

Method	Easy	Moderate	Hard
F-PointNet [22]	82.19	69.59	60.59
3D-CVF [44]	89.20	80.05	73.11
Focals Conv [3]	90.55	82.28	77.59
SFD [34]	**91.73**	84.76	77.92
TED-M [32]	91.61	**85.28**	80.68
Intervene3D	90.84	84.62	**81.22**

Table 3: Robustness to Conflicting Semantic Control on KITTI (Mod. AP_{3D}). BEV completion is our reimplementation.

Method	Normal	Conflict	ΔAP $\uparrow$	FP infl. $\downarrow$
BEV completion	82.90	74.85	-8.05	+20%
Prompt concat	88.36	83.66	-4.70	+12%
Intervene3D	**89.75**	**88.38**	**-1.37**	**+4%**

Table 4: Performance evaluation of state-of-the-art 3D detection approaches on the nuScenes validation benchmark.

Method	**Mod**	**2D Backbone**	**mAP**	**NDS**
MVP [43]	LC	-	66.4	70.5
FusionPainting [38]	LC	-	68.1	71.6
TransFusion [1]	LC	ResNet-50	67.5	71.3
DeepInteraction [41]	LC	ResNet-50	69.9	72.6
BEVFusion [10]	LC	SwinT	69.6	72.1
BEVFusion4D-S [2]	LC	SwinT	70.9	72.9
CMT [39]	LC	V2-99	70.3	72.9
EA-LSS [7]	LC	SwinT	71.2	73.1
FusionFormer [6]	LC	V2-99	70.0	73.2
SparseFusion [36]	LC	SwinT	71.0	73.1
CLS-3D [19]	LC	V2-99	71.6	74.3
SC3D [18]	LC	V2-99	71.8	74.5
BEVFormer++ [20]	LC	V2-99	71.8	74.5
Intervene3D	**LC**	**Convnext**	**72.1**	**74.9**

and Cyclist across all three difficulty levels. On nuScenes v1.0, we report mAP and NDS. Robustness is measured by severity-swept AUC, relative degradation, and false-positive inflation under prompt–evidence conflict.

Baselines and implementation. We report two types of comparisons: controlled in-code baselines and external published benchmarks. For controlled results, all variants use the same voxelization, image resolution, ConvNeXt image backbone, BEV decoder, detection head, optimizer, training schedule, and robustness protocol; only the tested module is changed. These variants include uni-modal detectors, BEVFusion-style fusion, TransFusion-style cross-modal attention, BEV completion, prompt concatenation/FiLM, and a parameter-matched residual refiner. External published methods in Tables 1, 2, and 4 are used only for benchmark context, since their backbones and training recipes may differ. LiDAR voxel size is 0.05 m on KITTI. The scene-description prompt encodes coarse count intervals, ego-centric spatial priors, and uncertainty cues. The prompt predictor Φ is pretrained, frozen during detector training, and detached before IaO. IaO uses $K{=}3$ refinement steps in all reported results. KITTI models are trained with AdamW (learning rate 2×10^{-4}, weight decay 10^{-2}, step decay at epochs 50 and 65) for 74 epochs with batch size 2 per GPU (Supplement Table 13). Code and configurations: `https://github.com/Constantino2019-et/Intervene3D`.

Control-signal construction. During evaluation, Intervene3D uses a predicted scene-description prompt as the control signal. Because external sources such as HD maps, V2X, route priors, or manual descriptions are unavailable, the prompt is generated only from allowed sensor-derived inputs and encodes coarse context such as presence cues, count ranges, ego-centric regions, scene layout, and reliability. Dataset annotations supervise the prompt predictor only during training. At validation/testing, no ground-truth boxes, counts, labels, or annotation-derived spatial priors are used; ground-truth prompts are reserved only for oracle upper-bound diagnostics.

Robustness Protocol. We apply test-time-only corruptions with fixed model parameters: LiDAR sparsification with keep ratios $\rho \in \{1.0, 0.75, 0.5, 0.25, 0.1\}$, camera–LiDAR extrinsic perturbations with $\sigma_R \in \{0, 1, 2, 3, 5\}^\circ$ and $\sigma_t \in \{0, 0.05, 0.1, 0.2, 0.3\}$ m, and image occlusion with $\gamma \in \{0, 0.1, 0.2, 0.3, 0.4\}$. Stochastic corruptions are averaged over three seeds. Robustness-AUC is the trapezoidal mean of Moderate AP_{3D} over the physical severity values (keep-ratio deficit $1-\rho$, rotation σ_R in degrees, and occlusion fraction γ, respectively), in AP points and not normalized by the clean AP: $\mathrm{AUC} = \frac{1}{s_{S-1}-s_0} \sum_{i=0}^{S-2} (s_{i+1} - s_i) \frac{P(s_i)+P(s_{i+1})}{2}$, where $P(s_i)$ is the Moderate AP_{3D} at severity s_i, s_0 is the uncorrupted setting, and $S = 5$ levels are used per corruption family.

4.2 Main Results

KITTI validation set. Table 1 shows that Intervene3D achieves strong LC detection performance across Vehicle, Pedestrian, and Cyclist categories. For Vehicle detection, Intervene3D obtains the best Moderate and Hard mAP among the compared LC methods, with 89.75 and 88.00, respectively. For Pedestrian detection, it consistently outperforms all reported LC baselines across Easy, Moderate, and

Table 5: Robustness to Sensor Corruptions on KITTI.

Method	Sparse	Extri. AUC ↑	Occl.
BEVFusion	78.20	70.15	88.40
TransFusion	80.75	72.60	88.95
BEV Completion	82.90	74.85	89.20
Prompt Concat. (Standard)	76.40	69.30	88.10
Residual Refiner	**83.95**	76.81	89.35
Intervene3D	83.83	**78.71**	**89.45**

Table 6: Prompt-source diagnostics on KITTI validation. ΔAP denotes conflict-induced AP change; FP denotes false-positive inflation.

Control	Mod. AP_3D ↑	Extri AUC ↑	ΔAP	FP ↓
Neutral	88.91	77.18	-2.42	10%
Random	88.47	76.44	-3.08	17%
Swapped	88.18	75.92	-3.56	22%
Predicted	89.75	78.71	-1.37	4%
Oracle coarse	**90.28**	**79.34**	$\mathbf{-0.86}$	**3%**

Table 7: Isolation study of spectral tokenization and ϕ correction under camera-LiDAR extrinsic perturbation. Phase correction only is applied before spatial-domain fusion.

Variant	Mod. AP_{3D} ↑	Extri. AUC ↑
Spatial fusion baseline	86.73	74.20
w/o Phase correction	88.13	75.50
Phase correction only	88.83	76.30
Phase-aware spectral attention	**89.75**	**78.71**

Table 8: Projected-latent feasibility audit. Normalized radius is the fraction of the permitted trust-region radius used.

Setting	Viol (%) ↓	Mean rad	Max. rad
Clean	0.00	0.842	0.998
LiDAR sparsification	0.00	0.889	0.999
Extrinsic perturbation	0.00	0.918	1.000
Image occlusion	0.00	0.934	1.000

Table 9: Controlled component ablation on KITTI validation. Higher AP and extrinsic AUC and lower conflict-induced FP inflation are preferred.

Configuration	Mod. AP_{3D} ↑	Extri. AUC ↑	FP ↓
Spatial fusion	86.73	74.20	15%
Spectral + Prompt	88.36	75.62	12%
Spectral + Residual	88.93	76.81	12%
IaO Only	89.28	77.82	9%
Spectral Only	89.56	77.40	8%
Full Intervene3D	**89.75**	**78.71**	**4%**

Table 10: Trust metric comparison under extrinsic perturbation and prompt–evidence conflict.

Metric	Extri. AUC ↑	FP ↓
Identity	76.58	8.9%
Fixed isotropic	77.21	6.8%
Learned diagonal	**78.71**	**4%**

Hard settings, achieving 69.37, 62.59, and 53.73 mAP. For Cyclist detection, Intervene3D achieves the best Easy and Hard results, while BEVFormer++ remains slightly higher on the Moderate setting. Compared with Intervene3D w/o IaO, the full model improves Pedestrian AP by 1.09, 2.80, and 0.35 percentage points under Easy, Moderate, and Hard settings, respectively, and Cyclist AP by 1.36, 0.26, and 1.67 points. For Vehicle detection, the full model improves Moderate and Hard AP by 0.19 and 0.08 points, respectively, while Easy AP decreases by 0.25 points. Overall, these results demonstrate that Intervene3D provides competitive LC detection performance on KITTI, with the largest improvement observed for Pedestrian Moderate detection.

KITTI test set. Table 2 shows that Intervene3D reaches 90.84, 84.62, and 81.22 AP_{3D} for Car detection under Easy, Moderate, and Hard settings. Although SFD achieves the highest Easy AP and TED-M achieves the highest Moderate AP, Intervene3D obtains the best Hard performance, exceeding TED-M by 0.54 percentage points. This is consistent with the motivation of the method, although the test-set margin is small and no confidence intervals are available.

nuScenes validation. Table 4 shows that Intervene3D achieves 72.1 mAP and 74.9 NDS, exceeding the strongest listed LC baseline by 0.3 mAP and 0.4 NDS. The improvement over recent strong methods is modest, but it shows that the proposed framework remains competitive on a larger-scale multi-class benchmark. However, because the compared methods use different 2D backbones and training configurations, these results should be interpreted as benchmark-level evidence of competitiveness rather than a fully isolated measurement of the proposed module.

Prompt-evidence conflict. Table 3 shows that Intervene3D achieves 89.75 Moderate AP_{3D} under normal control and 88.38 under conflicting control, corresponding to a decrease of 1.37 percentage points. In comparison, prompt concatenation and BEV completion experience larger decreases of 4.70 and 8.05 points, respectively. Intervene3D also records the lowest false-positive inflation of 4%, compared with 12% for prompt concatenation and 20% for BEV completion. These results indicate that the proposed metric-bounded intervention mechanism reduces sensitivity to conflicting control while maintaining stronger detection performance. However, the latent trust-region bound does not guarantee the correctness of decoded detections.

5 Ablation Studies

We evaluate spectral alignment, phase correction, structured control, and intervention-aware optimization (IaO) on KITTI validation. Unless stated otherwise, AP denotes Moderate AP_{3D}, and robustness AUC summarizes performance across corruption severities.

Component ablation. Table 9 shows that full Intervene3D achieves 89.75 AP, 78.71 extrinsic AUC, and 4% conflict-induced FP inflation. Removing the spectral branch (IaO only) or IaO (Spectral Only) reduces AP to 89.28 and 89.56, respectively. The matched residual refiner achieves 88.93 AP with 12% FP inflation. These results indicate complementary benefits from spectral fusion and bounded intervention. All component variants share the backbone, training schedule, and evaluation protocol.

Spectral alignment. Table 7 isolates tokenization and phase correction. Spectral tokens without phase correction improve AP from 86.73 to 88.13, while phase correction alone achieves 88.83. Their combination reaches 89.75 AP and 78.71 AUC, gaining 3.02 AP and 4.51 AUC points over spatial fusion. These results support the joint contribution of spectral representation and phase alignment. Both tables use the final full cross-modal attention model.

Corruption robustness. Table 5 reports robustness under LiDAR sparsification (applied in data preprocessing), extrinsic perturbation, and image occlusion. Intervene3D achieves the highest extrinsic (78.71) and occlusion (89.45) AUC, while its sparsification AUC (83.83) is slightly below the residual refiner (83.95), so Intervene3D leads in two of the three corruption families.

Control reliability. Table 6 compares neutral, random, swapped, predicted, and oracle control. Predicted control achieves 89.75 AP, 78.71 AUC, and 4% conflict-induced FP inflation, outperforming neutral, random, and swapped alternatives. Oracle control reaches 90.28 AP and 79.34 AUC, indicating remaining headroom. These comparisons do not fully exclude information leakage.

Latent feasibility. We measure violations of the trust-region bound $\|z_K - z_0\|^2_{H(x)} \leq 2\delta(u)$ using a numerical tolerance of 10^{-6}. Table 8 reports zero observed violations across all settings. Mean normalized radius increases from 0.842 (clean) to 0.934 (occlusion), confirming empirical latent feasibility without guaranteeing detection correctness.

Trust metric comparison. Table 10 compares identity, fixed-isotropic, and learned-diagonal metrics. The learned diagonal achieves the lowest FP inflation (4%) and the highest extrinsic AUC (78.71). These findings support uncertainty-adaptive intervention but do not imply detection-level safety guarantees.

Computational efficiency. Compared with BEVFusion, Intervene3D increases computation from 102 to 110 GFLOPs (7.8%), reduces throughput from 22 to 20 FPS, and increases GPU memory from 8.4 to 8.9 GB.

6 Discussion and Limitations

Intervene3D enables bounded test-time intervention while keeping sensor evidence fixed, but several limitations remain. **External and predicted control.** Our primary experiments use sensor-predicted control, which may encode dataset-specific priors. A preliminary check with a KITTI OXTS ego-motion prior (Supplement Sec. N) is compatible with external guidance, but generalization to HD maps and V2X remains unverified, and inaccurate or conflicting control may degrade detection. **Trust metric reliability.** IaO depends on a reliable trust metric and adaptive intervention budget; an inaccurate metric may suppress useful corrections or permit excessive refinement. **Latent-space guarantee.** Bounding latent movement does not guarantee detection correctness or prevent false positives after decoding and NMS; we therefore evaluate both latent feasibility and detection-level robustness. **False negatives and temporal extension.** IaO cannot reliably recover objects absent from both camera and LiDAR evidence.

Acknowledgments and Disclosure of Funding

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2025-02263031). This work was also partially supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2025-23524806) and by the Technology Innovation Program funded by the Korea government (MOTIR) (No. RS-2022-00154678).

References

[1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1090–1099, 2022.

[2] Hongxiang Cai, Zeyuan Zhang, Zhenyu Zhou, Ziyin Li, Wenbo Ding, and Jiuhua Zhao. Bevfusion4d: Learning lidar-camera fusion under bird's-eye-view via cross-modality guidance and temporal aggregation. *arXiv preprint arXiv:2303.17099*, 2023.

[3] Yukang Chen, Yanwei Li, Xiangyu Zhang, Jian Sun, and Jiaya Jia. Focal sparse convolutional networks for 3d object detection. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 5418–5427. IEEE, 2022.

[4] Bei Cheng, Bo Wang, Qingwang Wang, and Tao Shen. The spatio-temporal and frequency domain learning framework for moving object detection in satellite video. *IEEE Geoscience and Remote Sensing Letters*, 2025.

[5] Cheng-Yao Hong, Li-Heng Wang, and Tyng-Luh Liu. Promptable 3-d object localization with latent diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[6] Chunyong Hu, Hang Zheng, Kun Li, Jianyun Xu, Weibo Mao, Maochun Luo, Lingxuan Wang, Mingxia Chen, Qihao Peng, Kaixuan Liu, et al. Fusionformer: A multi-sensory fusion in bird's-eye-view and temporal consistent transformer for 3d object detection. *arXiv preprint arXiv:2309.05257*, 2023.

[7] Haotian Hu, Fanyi Wang, Jingwen Su, Yaonong Wang, Laifeng Hu, Weiye Fang, Jingwei Xu, and Zhiwang Zhang. Ea-lss: Edge-aware lift-splat-shot framework for 3d bev object detection. *arXiv preprint arXiv:2303.17895*, 2023.

[8] Yufeng Jin, Chongben Tao, Chen Wang, Feng Cao, Xizhao Luo, Zhen Gao, Zufeng Zhang, Sifa Zheng, and Yuan Zhu. Ffaformer: A feature fusion attention former of point-image for 3d object detection. *IEEE Transactions on Vehicular Technology*, 2026.

[9] Zhenpeng Li, Xiao Zhao, Jingwei Bian, Biao Liu, Wei Li, and Lihua Zhang. V-fusion: 2d detection-enhanced multimodal 3d bev object detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[10] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in neural information processing systems*, 35:10421–10434, 2022.

[11] Hongyi Lin, Shouqun Ming, Yang Liu, and Xiaobo Qu. Aefusion: An attention-based ensemble learning approach for bev fusion perception in autonomous modular buses. *IEEE Transactions on Intelligent Vehicles*, 2024.

[12] Yonglei Liu, Li Qin, Kun Hao, Zhisheng Li, Lu Zhao, and Xiaofang Zhao. Sfd-iafnet: 3d detection method for vehicle small objects based on multi-scale feature enhancement and cross-modal interlaced attention. *Measurement Science and Technology*, 36(9):095406, 2025.

[13] Yuhao Liu, Bing Tu, Bo Liu, Yan He, Jun Li, and Antonio Plaza. Spatial frequency domain transformation for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[14] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022.

[15] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023.

[16] Heitor R Medeiros, Atif Belal, Srikanth Muralidharan, Eric Granger, and Marco Pedersoli. Visual modality prompt for adapting vision-language object detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2172–2182, 2025.

[17] Husnain Mushtaq, Xiaoheng Deng, Ping Jiang, Shaohua Wan, Mubashir Ali, and Irshad Ullah. Gfa-smt: Geometric feature aggregation and self-attention in a multi-head transformer for 3d object detection in autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 26(3):3557–3573, 2024.

[18] Husnain Mushtaq, Xiaoheng Deng, Roohallah Alizadehsani, Muhammad Shahid Iqbal, Tamoor Khan, and Adeel Ahmed Abbasi. Sc3d: Semantic-guided and class-adaptive cross-domain fusion for 3d object detection in autonomous vehicles. *Expert Systems with Applications*, 268: 126359, 2025.

[19] Husnain Mushtaq, Sohaib Latif, Muhammad Saad Bin Ilyas, Syed Muhammad Mohsin, and Mohammed Ali. Cls-3d: Content-wise lidar-camera fusion and slot reweighting transformer for 3d object detection in autonomous vehicles. *IEEE Access*, 13:69840–69856, 2025.

[20] Shazib Qayyum, Xiaoheng Deng, Husnain Mushtaq, Ping Jiang, Shaohua Wan, and Irsha Ullah. Bevformer++: Enhancing bev fusion with normalized embedding and range attention for 3d object detection. *Expert Systems with Applications*, page 131131, 2026.

[21] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[22] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 918–927, 2018.

[23] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[24] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10529–10538, 2020.

[25] Yongchao Song, Jiping Bi, Lijun Sun, Zhaowei Liu, Yahong Jiang, and Xuan Wang. Freq-3dlane: 3d lane detection from monocular images via frequency-aware feature fusion. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

[26] Zhonghan Tao, Qianqian Xin, Yunpu Liu, Yanan Li, and Xichang Liang. F3dnet: Frequency-domain fusion for 3d object detection. In *2025 7th International Conference on Frontier Technologies of Information and Computer (ICFTIC)*, pages 749–754. IEEE, 2025.

[27] Muhammad Uzair, Jian Dong, Ronghua Shi, Husnain Mushtaq, and Irshad Ullah. Channelwise and spatially guided multimodal feature fusion network for 3-d object detection in autonomous vehicles. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024.

[28] Muhammad Uzair, Jian Dong, Ronghua Shi, Husnain Mushtaq, and Irshad Ullah. Trans-lgs: Transformer-based local graph structure for 3d object detection in autonomous vehicles. *IEEE Transactions on Vehicular Technology*, 2025.

[29] Lingyun Wang, Bingjie Wang, Jay Chhablani, Jose Alain Sahel, and Shaohua Pi. Freqformer: Frequency-domain transformer for 3-d reconstruction and quantification of human retinal vasculature. *IEEE Transactions on Biomedical Engineering*, 2025.

[30] Xiaoyi Wang, Jun Du, Qingxuan Lv, Peng Wang, Xianchao Zhang, Jian Cheng, and Henry Leung. Hyperspectral anomaly detection using dual-branch network based on frequency domain learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

[31] Zhitao Wen, Jinhai Liu, Huaguang Zhang, and Fengyuan Zuo. Exploring fine-grained visual-text feature alignment with prompt tuning for domain-adaptive object detection. *IEEE Transactions on Cybernetics*, 2025.

[32] Hai Wu, Chenglu Wen, Wei Li, Xin Li, Ruigang Yang, and Cheng Wang. Transformation-equivariant 3d object detection for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2795–2802, 2023.

[33] Hai Wu, Chenglu Wen, Shaoshuai Shi, Xin Li, and Cheng Wang. Virtual sparse convolution for multimodal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21653–21662, 2023.

[34] Xiaopei Wu, Liang Peng, Honghui Yang, Liang Xie, Chenxi Huang, Chengqi Deng, Haifeng Liu, and Deng Cai. Sparse fuse dense: Towards high quality 3d detection with depth completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5418–5427, 2022.

[35] Xinhao Xiang, Simon Dräger, and Jiawei Zhang. 3diffusiondet: Diffusion model for 3d object detection with robust lidar-camera fusion. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10157–10161. IEEE, 2026.

[36] Yichen Xie, Chenfeng Xu, Marie-Julie Rakotosaona, Patrick Rim, Federico Tombari, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17591–17602, 2023.

[37] Mingwei Xing, Yao Wu, Yachao Zhang, and Yanyun Qu. Prompt-driven multi-modal unsupervised domain adaptation for 3d semantic segmentation. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2025.

[38] Shaoqing Xu, Dingfu Zhou, Jin Fang, Junbo Yin, Zhou Bin, and Liangjun Zhang. Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection. In *2021 IEEE international intelligent transportation systems conference (ITSC)*, pages 3047–3054. IEEE, 2021.

[39] Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer: Towards fast and robust 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 18268–18278, 2023.

[40] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

[41] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. Deepinteraction: 3d object detection via modality interaction. *Advances in Neural Information Processing Systems*, 35:1992–2005, 2022.

[42] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.

[43] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Multimodal virtual point 3d detection. *Advances in Neural Information Processing Systems*, 34:16494–16507, 2021.

[44] Jin Hyeok Yoo, Yecheol Kim, Jisong Kim, and Jun Won Choi. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XXVII 16*, pages 720–736. Springer, 2020.

[45] Can Zhang, Feipeng Da, and Shaoyan Gai. Point clouds feature frequency domain analysis based on multilayer perceptron. *The Visual Computer*, 41(2):1007–1020, 2025.

[46] Hanxue Zhang, Zetong Yang, Yanan Sun, Li Chen, Fei Xia, Fatma Güney, and Hongyang Li. Test-time correction: An online 3d detection system via visual prompting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[47] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018.

[48] Yin Zhou, Pei Sun, Yu Zhang, Dragomir Anguelov, Jiyang Gao, Tom Ouyang, James Guo, Jiquan Ngiam, and Vijay Vasudevan. End-to-end multi-view fusion for 3d object detection in lidar point clouds. In *Conference on Robot Learning*, pages 923–932. PMLR, 2020.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made match the experimental results.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations have been discussed.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have discussed the details fully and the code will be attached in the supplementary.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code will be attached in the supplementary.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have included all the details

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have included statistical significance.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have included the information.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed the broader impacts

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have properly cited the used code, data and models, etc

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

A Supplementary Material Overview

This supplementary material provides implementation and evaluation details for Intervene3D, a controlled-inference framework for multimodal BEV perception under partial observability. We describe the distinction between controlled and external benchmark comparisons, the construction of sensor-predicted control, the metric-bounded intervention operator (IaO), its theoretical scope, phase-aware spectral alignment, robustness and ablation results, implementation settings, and limitations. We also separate the proposed KITTI OXTS subset analysis from the full-validation evaluation because the subset protocol and numerical results require independent verification.

The primary evidence for the proposed modules comes from controlled comparisons using matched detector components and evaluation conditions. Published KITTI and nuScenes comparisons provide benchmark context but do not isolate the effect of the proposed modules when architectures, training schedules, input resolutions, or temporal settings differ.

B Controlled Evaluation Protocol

Controlled configurations share LiDAR voxelization, image resolution, image backbone, BEV decoder, detection head, optimizer, augmentation policy, training schedule, and corruption protocol; only the component under investigation is changed. Comparisons cover spatial fusion, spectral fusion with and without phase correction, prompt concatenation/FiLM, BEV completion, a parameter-matched residual refiner, and IaO variants. The residual refiner assesses whether unbounded feature updates yield comparable accuracy, robustness, and false-positive behavior to bounded intervention.

External benchmark results on KITTI and nuScenes are reported separately. Cross-paper improvements should not be attributed solely to IaO or spectral alignment because the published systems may differ in backbone, detection head, input resolution, sensor utilization, or temporal context.

C Control Signal Construction and Non-Leakage

Intervene3D represents structured control by τ. A control generator maps τ and the permitted scene context into constraint parameters $\mathcal{S}$ for the differentiable energy $C(z; \mathcal{S})$. We distinguish three sources.

Exogenous control. HD maps, route priors, V2X messages, weather metadata, and policy constraints are possible externally supplied sources. They are not used in the main KITTI and nuScenes evaluation.

Sensor-predicted control. The main results use $\tau = \Phi(x)$, where a lightweight predictor Φ receives permitted RGB–LiDAR information and outputs coarse class-presence cues, count intervals, ego-centric admissibility fields, layout descriptors, and reliability information. The predictor is trained using coarse targets constructed from training annotations, then frozen during detector optimization. Its output is detached before IaO and is computed once per input rather than being regenerated after refinement.

Oracle coarse control. Coarse summaries derived from evaluation annotations are used only in a labeled oracle upper-bound diagnostic, not in the main model results.

C.1 Evaluation-Time Information Flow

At validation and test time, no ground-truth boxes, object counts, class labels, or annotation-derived admissibility fields are passed to Φ or IaO. The annotation-based target-construction pathway is disabled. Sensor-predicted control is not the same as exogenous information: both z_0 and τ depend on the observed inputs, even though their computational pathways and optimization roles are separated.

C.2 Control-Source Diagnostics

The main-paper prompt-source study compares neutral, random, swapped, predicted, and oracle coarse control on KITTI validation. The reported results are reproduced in Table 11. The conflict

Table 11: Control-source diagnostics on KITTI validation (values from the main paper).

Control	AP ↑	Extri. AUC ↑	Conflict ΔAP ↑	FP ↓
Neutral	88.91	77.18	−2.42	10%
Random	88.47	76.44	−3.08	17%
Swapped	88.18	75.92	−3.56	22%
Predicted	89.75	78.71	−1.37	4%
Oracle coarse	**90.28**	**79.34**	**−0.86**	**3%**

metric $\Delta\mathrm{AP}$ is the difference between conflict and normal Moderate AP_{3D}; values closer to zero indicate lower degradation. False-positive (FP) inflation is reported as a relative percentage.

Random and swapped control yield lower performance than predicted control in the reported runs. These interventions probe reliance on scene-specific information but do not by themselves prove absence of information leakage or generalization to independent external guidance.

D Metric-Bounded Controlled Inference

For sensor observation x, the evidence encoder produces the initial BEV latent $z_0 = \phi(x)$. The model predicts diagonal log-variance $s(x)$ and uses the positive-definite trust metric

$$H(x) = \mathrm{diag}(\exp(-s(x))). \tag{4}$$

Given a fixed control specification $\mathcal{S}$, IaO seeks to reduce the differentiable constraint energy subject to a bounded deviation from the evidence latent:

$$\min_z C(z;\mathcal{S}) \quad \text{s.t.} \quad \|z - z_0\|_H^2 \le 2\delta(u), \tag{5}$$

where $\|v\|_H^2 = v^\top H v$ and

$$\delta(u) = \delta_{\min} + (\delta_{\max} - \delta_{\min})u^\gamma, \qquad u \in [0,1]. \tag{6}$$

A higher uncertainty score can therefore permit a larger metric-norm intervention budget. This is not a guarantee that the intervention will improve predictions.

D.1 Constraint Energy

For class c, let $L_c(z) \in \mathbb{R}^{H_b \times W_b}$ be the BEV logits and $P_c(z) = \sigma(L_c(z))$. With count interval $[m_c, M_c]$, soft occupancy mass $N_c(z) = \sum_{i,j} P_c(z)_{ij}$, admissibility field $A_c \in [0,1]^{H_b \times W_b}$, and nonnegative weights α_c, β_c, define

$$\mathcal{L}_c^{\mathrm{count}}(z) = \mathrm{ReLU}(m_c - N_c(z))^2 + \mathrm{ReLU}(N_c(z) - M_c)^2, \tag{7}$$

$$\mathcal{L}_c^{\mathrm{spatial}}(z) = \frac{1}{H_b W_b} \sum_{i,j} (1 - A_c(i,j)) P_c(z)_{ij}, \tag{8}$$

$$C(z;\mathcal{S}) = w(u) \sum_c \left[\alpha_c \mathcal{L}_c^{\mathrm{count}}(z) + \beta_c \mathcal{L}_c^{\mathrm{spatial}}(z)\right], \tag{9}$$

where $w(u) = \mathrm{clip}(u^\kappa, 0, 1)$. These differentiable surrogates are used for refinement; satisfying them does not guarantee correct object boxes or post-NMS counts.

D.2 Refinement Update and Feasibility

Starting from z_0, the update at step k is

$$g_k = \nabla_z C(z_k;\mathcal{S}), \tag{10}$$

$$\widetilde{z}_{k+1} = z_k - \eta_k H^{-1} g_k, \tag{11}$$

$$v_k = \widetilde{z}_{k+1} - z_0. \tag{12}$$

The implementation uses the metric-radial feasibility mapping

$$z_{k+1} = \begin{cases} \widetilde{z}_{k+1}, & v_k^\top H v_k \le 2\delta(u), \\ z_0 + \sqrt{\dfrac{2\delta(u)}{v_k^\top H v_k}}\, v_k, & \text{otherwise.} \end{cases} \tag{13}$$

This map is the metric projection onto the ellipsoidal trust region when distance is measured in the H-norm; it is not generally the Euclidean projection onto that ellipsoid.

Algorithm: IaO refinement.

1. Compute z_0, H, uncertainty u, and the fixed control specification $\mathcal{S}$ from the input.
2. Initialize $z \leftarrow z_0$.
3. For $k = 0, \dots, K-1$, compute $g = \nabla_z C(z; \mathcal{S})$, form $\widetilde{z} = z - \eta_k H^{-1} g$, and apply Eq. (13) to obtain the next z.
4. Decode z_K with the original BEV decoder and detection head.

The main configuration uses $K = 3$. Feasibility follows directly because every mapped iterate belongs to the region $\{z : \|z - z_0\|_H^2 \le 2\delta(u)\}$. Convergence to the global minimum of the nonconvex constraint energy is not guaranteed.

D.3 Scope of the Guarantee

The exact guarantee is

$$\|z_k - z_0\|_H^2 \le 2\delta(u), \qquad k = 0, \dots, K, \tag{14}$$

up to numerical precision. It restricts latent displacement and does not guarantee semantic-constraint satisfaction, detection correctness, collision avoidance, or bounded false positives after decoding and NMS.

E Theoretical Properties

E.1 Restricted Gaussian KL Interpretation

Consider only Gaussian distributions with identical covariance H^{-1}:

$$q(z) = \mathcal{N}(\mu, H^{-1}), \qquad p(z \mid x) = \mathcal{N}(z_0, H^{-1}). \tag{15}$$

Their KL divergence satisfies

$$\mathrm{KL}(q\|p) = \frac{1}{2}(\mu - z_0)^\top H (\mu - z_0). \tag{16}$$

Thus a feasible mean μ satisfies $\mathrm{KL}(q\|p) \le \delta(u)$ within this fixed-covariance surrogate family. The detector is not assumed to specify a calibrated probabilistic posterior.

E.2 Bound on Changes in a Smooth Risk Proxy

Let $R(z)$ be differentiable and L-smooth with respect to the H-norm in the considered neighborhood of z_0:

$$R(z) \le R(z_0) + \nabla R(z_0)^\top (z - z_0) + \frac{L}{2}\|z - z_0\|_H^2. \tag{17}$$

For any feasible z, generalized Cauchy–Schwarz yields

$$R(z) - R(z_0) \le \sqrt{2\delta(u)}\,\|\nabla R(z_0)\|_{H^{-1}} + L\delta(u). \tag{18}$$

This is an upper bound on changes in the chosen smooth proxy, not a proof that AP or real-world risk improves.

Table 12: Spectral-tokenization and phase-correction isolation on KITTI validation, consistent with the updated main-paper table.

Variant	Mod. AP_{3D} ↑	Extri. AUC ↑
Spatial fusion baseline	86.73	74.20
Spectral tokens w/o phase correction	88.13	75.50
Phase correction only	88.83	76.30
Phase-aware spectral attention	**89.75**	**78.71**

E.3 Sensitivity to Trust-Metric Error

Suppose the intended metric is $H \succ 0$ and an estimate is $\widehat{H} = H + E$. If

$$\|H^{-1/2} E H^{-1/2}\|_2 \leq \epsilon < 1, \tag{19}$$

then

$$(1-\epsilon)\|v\|_H^2 \leq \|v\|_{\widehat{H}}^2 \leq (1+\epsilon)\|v\|_H^2. \tag{20}$$

In particular, feasibility under $\widehat{H}$ implies only $\|v\|_H^2 \leq 2\delta(u)/(1-\epsilon)$ under H. The bound motivates uncertainty calibration but does not establish that the learned metric is correct.

F Phase-Aware Spectral Alignment

A translation in a common spatial coordinate system induces a frequency-dependent phase shift under the Fourier shift theorem. Camera–LiDAR extrinsic perturbation is more general: it may involve 3D rotation, translation, projection differences, incomplete field-of-view overlap, and occlusion. Accordingly, phase correction is used as a feature-level alignment bias, not as an exact extrinsic-calibration solver.

For spectral representations $F_I = A_I e^{j\varphi_I}$ and $F_L = A_L e^{j\varphi_L}$, the alignment module predicts

$$\Delta\varphi = f_\Delta(\mathrm{GAP}(A_I), \mathrm{GAP}(A_L)), \qquad \varphi_L^{\mathrm{corr}} = \varphi_L - \Delta\varphi. \tag{21}$$

Corrected spectral features are divided into radial frequency bands and fused through cross-modal attention. Because the modalities may originate from different spatial sampling geometries, any shared-coordinate interpretation must be restricted to the aligned feature representation rather than raw 2D camera and 3D LiDAR frequency grids.

F.1 Spectral Isolation Results

The updated main-paper isolation table reports the values in Table 12. The full configuration has the highest reported Moderate AP_{3D} and extrinsic AUC in this comparison, improving over spatial fusion by 3.02 and 4.51 points, respectively. These results describe the listed experimental configurations; the phase model's ability to recover physical extrinsic parameters has not been established.

G Evidence-Trust Metric and Feasibility

The trust metric governs the geometry of allowable intervention while the uncertainty-dependent scalar budget controls its overall magnitude. A meaningful metric is expected to restrict intervention around reliable features and allow more freedom where evidence is uncertain; this interpretation depends on how well the uncertainty estimates track actual errors.

The main-paper feasibility audit reports zero observed violations across clean, sparse-LiDAR, extrinsic-perturbation, and image-occlusion settings. Mean normalized radius values are 0.842, 0.889, 0.918, and 0.934, respectively; the maximum reported normalized radius is no greater than 1.000. The reported numerical violation test is $\|z_K - z_0\|_H^2 > 2\delta(u) + 10^{-6}$.

Table 13: Reported Intervene3D implementation parameters. Entries marked "verify" or "implementation-dependent" must be resolved from the final training configuration before release.

Hyperparameter	Value
Optimizer	AdamW
Learning rate	2×10^{-4}
Weight decay	10^{-2}
KITTI / nuScenes epochs	74 / 400
Batch size per GPU	8 (verify)
Gradient norm clipping	10 (verify)
KITTI split	3,712 / 3,769
KITTI classes	Car, Pedestrian, Cyclist
Point features	x, y, z, intensity
Number of point features	4 (verify config)
KITTI voxel size	$0.1\,\mathrm{m}$
nuScenes voxel size	$0.2\,\mathrm{m}$
Max points per voxel	5 (verify)
Max voxels train/test	16,000 / 40,000 (verify)
3D backbone	`InterveneVirConv8x` (verify)
Image backbone	ConvNeXt
Image pretrained	True (verify)
Image feature dimension	512 (verify)
Image transform	FFT
LiDAR transform	3D FFT / Fourier (verify choice)
Frequency bands B	4 (verify)
Token dimension d	128 (verify)
Phase alignment	Enabled
Spectral fusion	Band attention
Cross-attention heads	16 (verify)
Cross-attention dropout	0.1 (verify)

Hyperparameter	Value
Scene-description prompt	Enabled
Prompt predictor Φ	Pretrained, frozen
Control at test time	Sensor-predicted
Prompt dimension	512 (verify)
Maximum count $K_{\max}$	20 (verify)
Prompt interval mass	0.8 (verify)
Scene conditioning	Enabled
Scene feature dimension	256 (verify)
IaO steps K	3
IaO step size η	0.1 (verify)
Maximum step size	50 (verify definition)
Count loss weight α	0.1 (verify)
Spatial loss weight β	0.02 (verify)
Metric bound	Enabled
Uncertainty gating	Enabled
Backtracking	Enabled (verify)
Energy tolerance	10^{-4} (verify)
Detection loss weight	1.0
Spectral loss weight	0.02 (verify)
Prompt-stability weight	0.05 (verify)
Refinement-consistency weight	0.05 (verify)
BEV feature dimension	256 (verify)
2D BEV backbone size	128 (verify meaning)
Dense head	`AnchorHeadSingle` (verify)
ROI head	`TEDMHead` (verify)
NMS threshold	0.1 (verify)

The trust-calibration comparison reports ECE/NLL/uncertainty–error correlation of $0.043/0.221/0.84$ for the learned diagonal metric, with extrinsic AUC 78.71 and rounded FP inflation 4%. The fixed-isotropic configuration reports $0.089/0.314/0.63$, AUC 77.21, and FP inflation 6.8%. Calibration measures are not reported for the identity metric. These results are empirical diagnostics, not a safety certificate.

H Implementation and Reproducibility

Compute resources. The supplied draft reports two NVIDIA A6000 GPUs with 48 GB memory each, approximately 8 hours per KITTI run, and approximately 336 hours per nuScenes run. These resource measurements and the batch-size assumptions should be checked against final training logs before submission.

Controlled comparisons share the data pipeline, image resolution, backbone, decoder, detection head, training schedule, optimizer, and augmentation settings, except for explicitly stated component changes. KITTI uses the standard 3,712/3,769 training/validation split. On nuScenes v1.0, performance is measured using mAP and NDS. Unless otherwise stated, the stated configuration uses single-frame LiDAR and camera observations.

I Model Configuration Summary

The full system consists of RGB and LiDAR feature encoders, frequency-band tokenization, phase-aware attention fusion, a frozen scene-description predictor, and the metric-bounded IaO operator. The refined BEV latent is decoded with the same detection head used for controlled baselines. Table 14 summarizes the configuration at a system level.

Table 14: Compact configuration summary.

Component	Configuration
Modalities	LiDAR and RGB camera
Control source	Sensor-predicted scene description
Test annotations for main control	Not used
Image backbone	ConvNeXt
Image spectral encoding	FFT and band tokenization
LiDAR spectral encoding	3D FFT / Fourier (verify choice)
Spectral fusion	Phase-aware cross-modal attention
Prompt predictor	Pretrained, frozen Φ
IaO	Three metric-bounded refinement steps
Trust metric	Learned diagonal $H(x)$
Detector decoder and head	Matched to controlled baseline
Optimizer	AdamW
KITTI / nuScenes epochs	74 / 400

J Experimental Results and Analysis

J.1 Robustness Under Sensor Corruption

The main paper reports severity-swept robustness AUC on KITTI Moderate AP_{3D}. Under LiDAR sparsification, Intervene3D obtains 83.83, marginally below the residual-refiner baseline (83.95). Under camera–LiDAR extrinsic perturbations, Intervene3D obtains 78.71, compared with 76.95 for the residual refiner. Under image occlusion, it obtains 89.56, compared with 89.35 for the residual refiner. Thus, the proposed model leads in two of these three reported perturbation families, not all three.

J.2 Prompt–Evidence Conflict

Under conflicting control, the main-paper results for Intervene3D change from 89.75 to 88.38 Moderate AP_{3D}, giving $\Delta AP = -1.37$ and rounded FP inflation of 4%. The prompt-concatenation baseline reports a 4.70-point AP decrease and 31% FP inflation, whereas BEV completion reports a 3.50-point decrease and 18% FP inflation. Because clean baseline AP differs across some reported ablations, these results should be compared only within the corresponding controlled protocol.

J.3 nuScenes Benchmark Context

Intervene3D reports 72.1 mAP and 74.9 NDS on nuScenes validation, compared with the strongest listed external results of 71.8 mAP and 74.5 NDS. Different backbones and training pipelines prevent attributing these modest differences to a single component.

K Ablation Studies

Component isolation. In the main-paper controlled ablation, spatial fusion obtains 86.73 Moderate AP_{3D} and 74.20 extrinsic AUC. Spectral fusion without IaO obtains 89.56 and 78.05; IaO without spectral fusion obtains 89.28 and 77.82; the full model obtains 89.75 and 78.71, with 4% conflict FP inflation. These comparisons suggest complementary contributions, although the incremental improvements differ by metric and setting.

Spectral tokenization and phase correction. The phase-isolation results are reproduced in Table 12; the combined variant obtains the largest reported AP and extrinsic AUC in that set. The number of frequency bands is reported as $B = 4$ in the configuration. The original supplementary

draft describes a sweep over $B \in \{2, 4, 8, 16\}$ but does not provide the corresponding numerical observations; consequently, an optimality claim is deferred until those measurements are supplied.

Structured control and unbounded refinement. Prompt concatenation/FiLM and a matched residual refiner assess whether gains can be achieved by standard conditioning or refinement without the proposed metric bound. The main-paper component table reports 88.36 AP, 75.62 extrinsic AUC, and 12% FP inflation for prompt concatenation/FiLM; the matched residual refiner reports 88.93, 76.81, and 9%, respectively.

Auxiliary losses. The detector objective combines detection, spectral consistency, prompt stability, and refinement consistency terms. The provided materials specify their form and example weights but do not include a numeric loss-removal table. No quantitative gain is attributed to an individual auxiliary loss here without those results.

L Sensitivity, Stability, and Efficiency

The extra IaO cost depends on the number of refinement steps K and the cost $C_{\mathcal{C}}$ of evaluating the constraint gradient, approximately $\mathcal{O}(KC_{\mathcal{C}})$ beyond the evidence encoder. Since H is diagonal, both preconditioning and the radial feasibility map require operations linear in latent dimension. Feasibility does not ensure monotonic energy reduction with arbitrary step sizes or monotonic improvement in AP.

The main paper reports an increase from 102 to 110 GFLOPs (approximately 7.8%), throughput declining from 22 to 20 frames/s, and memory increasing from 8.4 to 8.9 GB relative to the named BEVFusion baseline. Hardware, batch size, precision, and profiling protocol should accompany these numbers when the final experiment logs are available.

M Qualitative and Failure-Case Analysis

Qualitative analysis should contrast the original and refined detections under sparse LiDAR, camera–LiDAR misalignment, image occlusion, and conflicting control. Examples should include both corrected detections and failure cases. No qualitative figures were included in the supplied supplementary source, so no specific visual outcome is claimed here.

Expected failure modes include objects absent from both modalities, severe calibration drift, inaccurate or stale control, distribution shift, and nonlinear decoding that amplifies modest latent changes into false detections. The method changes an evidence-derived latent; it is not a complete missing-object reconstruction system.

N External Control and Test-Time Correction

The original supplementary draft contains a proposed KITTI OXTS subset analysis and a comparison with test-time correction. Unlike the main benchmark results, the supplied numbers have not been associated with a verified sample-ID list, subset size, or final evaluation logs. The tables below therefore preserve the supplied figures only as *unverified draft values*; they must not be presented as confirmed experimental findings.

N.1 KITTI OXTS Evaluation Protocol

The intended evaluation uses KITTI samples for which RGB, LiDAR, annotations, and OXTS measurements are jointly available. Before publication, report the exact subset size, selection rules, sample IDs, coordinate conversions, and whether all compared approaches were evaluated on identical examples. None of these details is established by the supplied text.

N.2 Exogenous Control from OXTS

The proposed external control is derived from GPS/IMU ego-motion trajectories and is intended to provide a coarse driving-corridor and heading prior. It replaces only the spatial portion of the

Table 15: KITTI OXTS subset control-source comparison.

Control source	AP_{3D} ↑	Extr. AUC ↑	Conflict ΔAP ↑	FP ↓
No control	88.01	74.20	–	–
Sensor-predicted	89.75	78.71	−1.12	4.2%
Exogenous OXTS	89.83	79.18	−0.87	3.8%
Oracle coarse	90.02	79.62	−0.61	3.4%

Table 16: Proposed test-time correction comparison.

Method	AP_{3D} ↑	LiDAR AUC ↑	Extr. AUC ↑	ΔAP ↑	FP ↓
Baseline detector	88.01	84.92	74.20	–	–
Test-time correction	89.18	86.96	77.63	−1.71	6.9%
Residual refiner	89.08	86.44	77.12	−1.86	7.9%
Intervene3D	**89.75**	**87.61**	**78.71**	−1.20	**4.2%**

structured control; the remaining control components and detector are unchanged. OXTS ego-motion is not equivalent to object-level semantics or independent ground-truth object annotations.

The OXTS-subset values must not be merged with the full KITTI prompt-source table. In particular, its sensor-predicted conflict change (-1.12) and FP inflation (4.2%) differ from the full-validation figures (-1.37 and 4%). Such differences require distinct subset/protocol documentation and cannot be explained by rounding alone.

N.3 Comparison with Test-Time Correction

The original draft also proposes a comparison with the test-time correction method of Zhang et al. [46] and a matched residual refiner. The common backbone, training schedule, corruption settings, and sample subset must be verified before making controlled-comparison claims.

These draft values suggest a potential advantage for Intervene3D within the proposed comparison; they do not establish it until the experimental protocol is confirmed. The -1.20 conflict change here differs from both the OXTS control-source table (-1.12) and the full-validation table (-1.37), and must be resolved from actual logs rather than harmonized by assumption.

N.4 Scope of External-Control Evidence

Even with a verified OXTS subset, the result would demonstrate only that IaO can consume a specific coarse ego-motion-derived spatial prior. It would not demonstrate generalization to arbitrary HD maps, V2X communications, route-planner constraints, or unconstrained external instructions.

O Limitations, Safety Considerations, and Broader Impact

Absent evidence and false negatives. IaO cannot reliably recover objects that are unobserved by both LiDAR and camera. It is intended to refine ambiguous evidence, not infer arbitrary missing scenes.

Predicted control and leakage risk. Because the main control is predicted from sensor observations and trained with coarse annotation supervision, it may inherit dataset priors. Neutral, swapped, random, and oracle comparisons help diagnose behavior but do not completely exclude leakage or distribution-shift failures.

Latent versus detection-level guarantees. The feasibility result constrains $\|z_K - z_0\|_H$ for the learned metric; it does not guarantee correct decoded boxes, valid post-NMS constraints, controlled false negatives, or safe downstream behavior.

Phase alignment and metric calibration. Phase correction is most defensible for mild displacement in a shared feature representation, not general 3D extrinsic recalibration. An inaccurate uncertainty metric may allow excessive intervention or suppress useful correction.

Deployment and societal impact. Automated driving is safety-critical. Before real-world deployment, the system requires evaluation under hardware faults, adverse weather, heavy occlusion, severe calibration drift, domain shift, and stale or adversarial control. A bounded latent change should not be represented as certified driving safety.

P Supplementary Conclusion

Intervene3D introduces an explicit and auditable interface for bounded test-time refinement of multimodal BEV latents. The supplied experiments report competitive clean detection, stronger robustness in two of three evaluated corruption families, reduced conflict-induced degradation, and zero observed violations of the implemented latent trust region. These observations support controlled latent intervention but do not certify detection correctness, general external-control transfer, or safety. Future release materials should resolve the unverified OXTS subset, test-time correction comparison, detailed model settings, and missing sensitivity measurements before those claims are treated as reproducible evidence.