TEMPCALIB: DIAGNOSING AND RECOVERING EXTRINSIC DRIFT WITH A FROZEN 3D DETECTOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-camera 3D detectors rely on accurate extrinsics to align cross-view observations, yet post-deployment drift can degrade downstream perception. Existing solutions either use dedicated online calibration or improve robustness without explicitly restoring corrupted extrinsics. The former adds a separate estimation procedure, while the latter leaves geometric errors unresolved. We investigate whether the deployed detector itself contains evidence for diagnosis and recovery. We propose TempCalib, a framework for unknown-camera fault diagnosis, localization, and rotational recovery using a frozen multi-camera 3D detector. It extracts reusable detector-derived evidence, detects persistent cross-view inconsistency on a camera graph, and generates target-excluded rotational proposals after localization. Cross-view feature consistency and fixed-correspondence epipolar geometry refine these proposals, and final states are compared on common visible support. Fixed clean references calibrate these signals; neither additional training nor per-candidate backbone or decoder passes are required. On a nuScenes-based benchmark, TempCalib detects and localizes every injected fault across three splits sharing a 96-condition single-camera rotational perturbation grid. On blind Test-A, it achieves a mean rotational residual of 0.988° under the matched 40-sequence setting, reduced to 0.894° with the extended recovery budget. On the two validation splits, the L4 configuration restores approximately 91–97% of the condition-level mAP/NDS loss. These results show that a frozen detector can provide evidence for explicit geometric recovery and restoration of its own perception.

1 INTRODUCTION

Camera calibration is part of a multi-view detector's inference mechanism, rather than merely a preprocessing step. Camera extrinsics determine where 3D queries access image features and how positional information aligns views (Wang et al., 2022b; Liu et al., 2022; 2023; Li et al., 2022). After deployment, mechanical or environmental changes may cause camera orientations to drift. The images can remain visually normal while incorrect geometry degrades 3D perception. This poses both a recovery and a diagnostic challenge: before correcting the geometry, the system must detect the fault and identify the camera.

A controlled StreamPETR study shows that 1°, 2°, and 5° rotations reduce mAP/NDS by 0.22/0.20, 0.95/0.76, and 4.08/3.03 points; 2–10 cm translations change either metric by less than 0.007 points over the evaluated range (Appendix B.8). We focus on post-deployment rotational extrinsic drift, its diagnosis and recovery, and downstream effects.

Existing approaches to calibration drift broadly follow two routes. Dedicated online calibration estimates sensor geometry through a separate deployment procedure (Dexheimer et al., 2022; Pätzold et al., 2022; Moravec et al., 2026). Geometry-aware or calibration-independent perception reduces dependence on precise camera geometry through representations or training (Xiong et al., 2023; Chen et al., 2023; Jiang et al., 2022), without necessarily correcting corrupted extrinsics. Online self-calibration, learned miscalibration detection, and calibration without additional training address related estimation or monitoring tasks (Kanai et al., 2023; Tahiraj et al., 2025; Luo et al., 2023). What remains unclear is whether an already deployed perception model can itself provide evidence for both fault diagnosis and explicit geometric recovery, without an additional calibration network.

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.

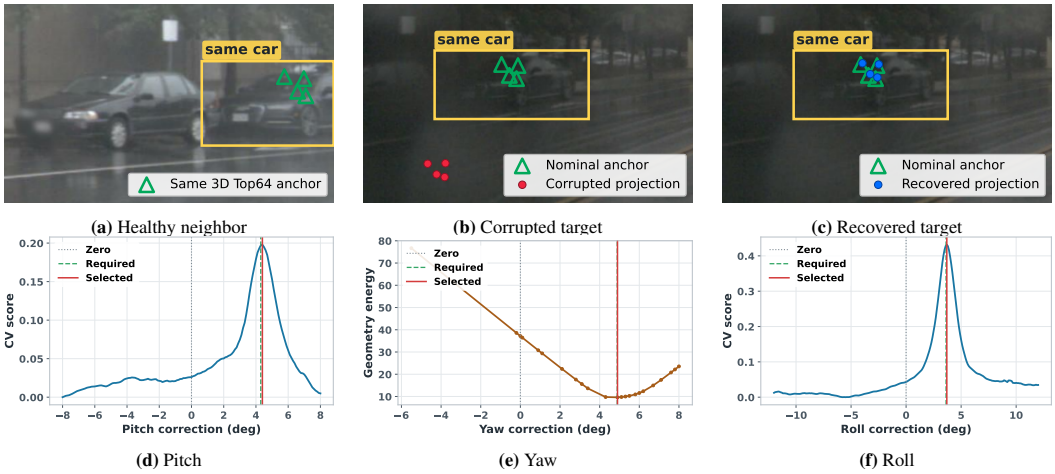


Figure 1: **Detector-derived evidence for rotational recovery.** Top: the same anchors, extracted once under corrupted calibration, align in the healthy neighbor but shift in the target under corrupted extrinsics (red) and realign after correction (blue); green marks nominal projections. Bottom: cross-view consistency peaks for pitch/roll and epipolar cost reaches a yaw minimum. Gray dotted, green dashed, and red solid lines denote zero, ground-truth, and selected corrections. Nominal target projections and ground truth are visualization-only references.

Our key insight is that extrinsic drift leaves structured inconsistencies in the same detector representations used for perception. A detector-derived 3D center induces sampling locations across cameras; when one camera drifts, its sampled evidence can become less consistent with the remaining views. Figure 1 illustrates this in cross-view projections and correction-sensitive responses along all three rotation axes. The detector is therefore not merely a calibration consumer but also a potential evidence source. We formulate **Perception-Driven Extrinsic Drift Diagnosis and Recovery**: a frozen detector must detect a fault, identify the affected camera, recover its rotation, and restore downstream perception.

Translating this idea into a recovery procedure requires separating comparison evidence from the geometry under evaluation. First, scene-wide appearance changes may alter consistency throughout the rig, whereas a single-camera fault should preferentially disturb its incident view pairs. Second, evidence incorporating the faulty camera can be biased by its geometry; recomputing it for each candidate also confounds comparison. Third, coupled rotations and candidate-dependent visibility can produce competing plausible states.

TempCalib addresses these issues with detector-derived evidence at multiple levels. Persistent camera-graph contrast detects and localizes camera-specific inconsistency. After localization, target-excluded evidence produces coarse rotational proposals. Cross-view feature consistency refines pitch and roll, while fixed-correspondence epipolar geometry refines yaw. Two recovery states are compared on common visible support to select the final correction. The detector stays frozen; visual features and decoder outputs are reused without per-candidate backbone or decoder passes.

Our contributions are threefold. **First**, we formulate joint post-deployment diagnosis and recovery using the affected frozen detector as the calibration evidence source. **Second**, we develop TempCalib, combining persistent graph diagnosis, target-excluded proposals, complementary feature and epipolar refinement, and common-support selection without additional training or per-candidate detector passes. **Third**, we construct a nuScenes-based benchmark linking unknown-camera diagnosis and rotational recovery to restoration of the same detector. Across three splits sharing 96 perturbations, TempCalib localizes every injected fault and substantially restores mAP/NDS performance.

2 Related Work

Geometry-aware multi-camera 3D perception. Multi-camera 3D perception uses camera geometry to associate evidence across views. LSS, BEVDet, and BEVDepth lift image features into

3D or BEV space (Philion & Fidler, 2020; Huang et al., 2021; Li et al., 2023); BEVFormer uses spatiotemporal attention (Li et al., 2022). DETR3D and PETR use projected queries or positional representations; StreamPETR further propagates object states over time (Wang et al., 2022b; Liu et al., 2022; Wang et al., 2023). Other studies improve geometric representations or corruption robustness (Xiong et al., 2023; Chen et al., 2023; Xie et al., 2023), while CFT learns calibration-free view transformations (Jiang et al., 2022). We instead diagnose and recover geometry using frozen-detector evidence.

Online extrinsic calibration and miscalibration recovery. Classical relative-pose estimation recovers camera geometry from correspondences (Lowe, 2004; Nistér, 2004; Hartley & Zisserman, 2004). Learned matching includes sparse attention-based methods (Sarlin et al., 2020; Lindenberger et al., 2023) and detector-free or dense approaches (Sun et al., 2021; Edstedt et al., 2024b). Online target-free calibration uses informative observations (Dexheimer et al., 2022), detected person keypoints (Pätzold et al., 2022), or essential-matrix tracking (Moravec et al., 2026); FloSR learns relative stereo rotation through flow-guided rectification (Kumar et al., 2024). Self-supervised methods jointly infer extrinsics, depth, and ego-motion (Kanai et al., 2023), while miscalibration monitoring detects sensor-geometry faults (Tahiraj et al., 2025). CalibNet, LCCNet, and BEVCalib instead learn LiDAR–camera alignment from multimodal inputs (Iyer et al., 2018; Lv et al., 2021; Yuan et al., 2025). These methods use dedicated estimators or address only part of diagnosis and recovery; TempCalib instead reuses the deployed detector's evidence for fault detection, camera localization, and rotational recovery.

Test-time adaptation and robustness. Test-Time Training, Tent, and CoTTA adapt models under distribution shift (Sun et al., 2020; Wang et al., 2021; 2022a); EATA and SAR address forgetting and stability (Niu et al., 2022; 2023). TempCalib instead freezes the perception model and corrects camera geometry rather than adapting to misaligned inputs.

3 METHOD

Problem setting. An *observation* i consists of time-aligned images captured by the camera rig, and a *sequence* m comprises observations indexed by $\mathcal{I}_m$. An *event* groups ordered sequences sharing one calibration state, without requiring continuous driving. For cameras $c \in \mathcal{C} = \{1, \dots, C\}$, the detector uses intrinsics $\mathbf{K}_{ic}$ and the current, potentially drifted, reference-to-camera extrinsics $\mathbf{T}^0_{ic} \in \mathrm{SE}(3)$, written compactly as $[\mathbf{R}^0_{ic} \mid \mathbf{t}^0_{ic}]$. Each event is clean or has a fixed rotational drift in one unknown camera $c^\star$, preserving its reference-frame center and leaving other camera calibrations unchanged. Given images, current calibration, a frozen detector, and fixed clean commissioning references, we predict a fault state $\hat{y}$ and, if faulty, a camera $t = \hat{c}$ and roll–pitch–yaw correction $\mathbf{c} = (c_r, c_p, c_y)^\top$. No evaluated event's faulty-camera identity, clean extrinsic, or injected perturbation is supplied. Corrections act by

$$\mathbf{T}_{it}(\mathbf{c}) = [\Delta\mathbf{R}(\mathbf{c})\mathbf{R}^0_{it} \mid \Delta\mathbf{R}(\mathbf{c})\mathbf{t}^0_{it}], \qquad \Delta\mathbf{R}(\mathbf{c}) = \mathbf{R}_x(c_r)\mathbf{R}_y(c_p)\mathbf{R}_z(c_y). \tag{1}$$

Figure 2 summarizes evidence collection, graph diagnosis, and rotational recovery. Shared evidence is collected before fault localization, whereas target-excluded evidence is constructed only after the affected camera has been identified. Candidate evaluation modifies only the target-camera geometry and its associated representations, without rerunning the visual encoder or decoder. Numerical settings are specified in Section 4.1.

3.1 STAGE 0: DETECTOR-DERIVED EVIDENCE COLLECTION

One visual encoding supplies backbone features $\mathbf{F}^{\ell_B}_{ic}$ and neck features $\mathbf{F}^{N}_{ic}$. The Shared-Evidence Decoder uses base queries without temporal memory. Final-layer confidence selects the top-$\mathcal{N}$ query indices $\mathcal{Q}^{\mathrm{all}}_i$, while layer ℓ_A supplies reference-frame centers and hidden states, defining $\mathcal{A}^{\mathrm{all}}_i = \{(\mathbf{x}^{\mathrm{all}}_{iq}, \mathbf{h}^{\mathrm{all}}_{iq}, \alpha^{\mathrm{all}}_{iq}) : q \in \mathcal{Q}^{\mathrm{all}}_i\}$. Here α is the final-layer confidence, separating query selection from intermediate geometric readout. We project each center onto the backbone grid using $(\mathbf{K}_{ic}, \mathbf{T}^0_{ic})$, bilinearly sample channel-wise L2-normalized $\mathbf{F}^{\ell_B}_{ic}$, and L2-normalize the sampled descriptor $\mathbf{d}_{icq}$. Validity ν_{icq} requires positive depth, finite in-bounds coordinates, and a finite descriptor.

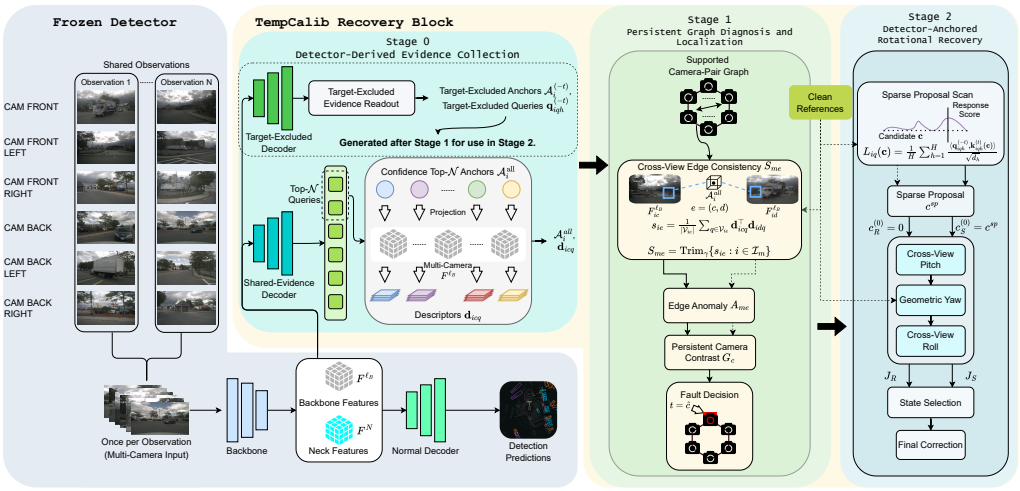


Figure 2: **The decoder as both a consumer of calibration and a source of calibration evidence.** All decoder paths share frozen parameters. The Normal Decoder performs detection; the Shared-Evidence Decoder supports diagnosis and recovery; the Target-Excluded Decoder constructs recovery evidence after localization. Dashed connections indicate components that use frozen clean references.

3.2 Stage 1: Persistent Graph Diagnosis and Localization

Commissioning evaluates all camera pairs $\mathcal{E}_0 = \{(c,d) : c,d \in \mathcal{C},\ c < d\}$ and freezes sufficiently supported edges $\mathcal{E}$, with clean sequence-level means μ_e and sample standard deviations σ_e. For $e = (c,d)$, jointly valid anchors $\mathcal{V}_{ie} = \{q \in \mathcal{Q}_i^{\text{all}} : \nu_{icq}\nu_{idq} = 1\}$ give $s_{ie} = \text{mean}_{q\in\mathcal{V}_{ie}} \mathbf{d}_{icq}^\top \mathbf{d}_{idq}$. We aggregate $S_{me} = \text{Trim}_\gamma\{s_{ie} : i \in \mathcal{I}_m\}$, where Trim_γ discards fraction γ from each tail before averaging; empty or non-finite inputs are omitted. Edge anomalies and camera contrasts are

$$A_{me} = \left[\frac{\mu_e - S_{me}}{\max(\sigma_e, \epsilon)}\right]_+, \qquad g_{mc} = \text{mean}_{e\in\mathcal{E}_{\text{inc}}(c)} A_{me} - \text{mean}_{e\in\mathcal{E}\setminus\mathcal{E}_{\text{inc}}(c)} A_{me}, \tag{2}$$

where $[x]_+ = \max(x, 0)$ and $\mathcal{E}_{\text{inc}}(c)$ contains edges incident to c. The contrast emphasizes camera-specific rather than rig-wide inconsistency.

The first M_D sequences form N_B ordered, equal-sized blocks $\mathcal{B}_\ell$. Define $B_{\ell c} = \text{Trim}_\gamma\{g_{mc} : m \in \mathcal{B}_\ell\}$, $G_c = \text{median}_\ell B_{\ell c}$, $\hat{c} = \arg\max_c G_c$, and $\hat{c}_\ell = \arg\max_c B_{\ell c}$. To require persistence across blocks, we declare

$$\hat{y} = \text{FAULT} \quad \Longleftrightarrow \quad G_{\hat{c}} > \tau \ \wedge \ \sum_{\ell=1}^{N_B} \mathbf{1}[\hat{c}_\ell = \hat{c}] \geq \eta_{\text{pers}}. \tag{3}$$

The threshold τ is frozen from clean reference events (Section 4.1). A fault is localized to $t = \hat{c}$; otherwise, $\hat{y} = \text{NO_FAULT}$ and recovery is skipped.

3.3 Stage 2: Detector-Anchored Rotational Recovery

Recovery uses $M_{\text{rec}} \geq M_D$ sequences, reusing existing observations when extending the budget. Each axis $a \in \{r, p, y\}$ has search set $\Omega_a = \{-r_a, -r_a + \delta_a, \ldots, r_a\}$, with radius r_a and step δ_a. The notation $\mathbf{c}[a \leftarrow v]$ replaces the total correction on axis a, not an incremental rotation.

Target-excluded sparse proposal. After localization, the Target-Excluded Decoder reuses neck features in a non-temporal pass per observation, excluding camera t from feature, region-selection, and token paths. The same layer readout and confidence selection produce a separate query set

$\mathcal{Q}_i^{(-t)}$; retaining $\mathcal{Q}_i^{\mathrm{sp}} = \{q \in \mathcal{Q}_i^{(-t)} : \nu_{itq}^N(\mathbf{0}) = 1\}$ fixes its anchor set $\mathcal{A}_i^{(-t)}$. Here $\nu_{itq}^N(\mathbf{c})$ denotes target-neck projection validity under correction $\mathbf{c}$. Target exclusion isolates the reference queries from the faulty camera's input. The next layer's frozen query projection maps the layer-ℓ_A hidden states and query-position embeddings to fixed per-head queries $\mathbf{q}_{iqh}^{(-t)}$.

For candidate $\mathbf{c}$, positional encoding, spatial alignment, and gating rebuild the target key-input map from $\mathbf{F}_{it}^N$. Sampling at the reprojected $\mathbf{x}_{iq}^{(-t)}$ and applying the frozen key projection yield $\mathbf{k}_{iqh}^{(t)}(\mathbf{c})$. With H heads of dimension d_h,

$$
\begin{aligned}
L_{iq}(\mathbf{c}) &= \frac{1}{H} \sum_{h=1}^{H} \frac{\langle \mathbf{q}_{iqh}^{(-t)}, \mathbf{k}_{iqh}^{(t)}(\mathbf{c}) \rangle}{\sqrt{d_h}}, \\
S_i^{\mathrm{sp}}(\mathbf{c}) &= \frac{1}{|\mathcal{Q}_i^{\mathrm{sp}}|} \sum_{q \in \mathcal{Q}_i^{\mathrm{sp}}} \begin{cases} L_{iq}(\mathbf{c}), & \nu_{itq}^N(\mathbf{c}) = 1, \\ -\lambda, & \text{otherwise.} \end{cases}
\end{aligned}
\tag{4}
$$

L_{iq} averages pre-softmax attention logits. For nonempty $\mathcal{Q}_i^{\mathrm{sp}}$, all candidates share its support; invalid projections receive $-\lambda$ rather than being discarded.

Fixed temporal sampling aggregates scores at $\mathbf{c}_a(v) = \mathbf{0}[a \leftarrow v]$ into sequence curves $C_{m,a}(v)$ (Appendix A.3). We rank candidate corrections $v \in \Omega_a$ within each sequence using normalized ascending zero-based ordinal ranks, average them across sequences, and align the resulting curve with the frozen camera-specific clean reference:

$$
\begin{aligned}
R_{\mathrm{tar},a}(v) &= \frac{1}{M_{\mathrm{rec}}} \sum_{m=1}^{M_{\mathrm{rec}}} \frac{\mathrm{rank}_0(C_{m,a}(v))}{|\Omega_a| - 1}, \\
c_a^{\mathrm{sp}} &= \arg\max_{\delta \in \Omega_{\mathrm{NCC}}} \mathrm{NCC}\big(R_{\mathrm{tar},a}(v), R_{\mathrm{tar},a}^{\mathrm{clean}}(v - \delta)\big).
\end{aligned}
\tag{5}
$$

Normalized cross-correlation (NCC) uses overlapping valid points; its maximizing shift, rather than a raw-score peak, supplies the correction. The axis estimates form $\mathbf{c}^{\mathrm{sp}}$. Yaw candidates are $\mathcal{H}_y = \{0, c_y^{\mathrm{sp}}\} \cup \mathrm{Peak}_{K_P}(R_{\mathrm{tar},y})$, retaining up to K_P strongest local-maximum angles in addition to zero and the sparse estimate.

Dual-state refinement. To reduce dependence on the coarse proposal, branches $b \in \{R, S\}$ start from $\mathbf{c}_R^{(0)} = \mathbf{0}$ and $\mathbf{c}_S^{(0)} = \mathbf{c}^{\mathrm{sp}}$ and follow pitch $\rightarrow$ yaw $\rightarrow$ roll. Each update holds the other components of the current state $\mathbf{c}_b$ fixed.

Cross-view pitch and roll refinement. We reuse shared anchors $\mathcal{A}_i^{\mathrm{all}}$ and backbone features, with sources $\mathcal{S}_t = \mathcal{C} \setminus \{t\}$. Candidate reprojection changes target descriptors and validity, while source quantities remain fixed. For a binary mask W of valid pairs, define

$$
\begin{aligned}
\rho_{iqs}(\mathbf{c}) &= \mathbf{d}_{itq}(\mathbf{c})^\top \mathbf{d}_{isq}, \qquad W_{iqs}(\mathbf{c}) = \nu_{itq}(\mathbf{c}) \nu_{isq}, \\
\bar{\rho}_{iq}(\mathbf{c}; W) &= \mathrm{mean}_{s \in \mathcal{S}_t :\, W_{iqs} = 1} \, \rho_{iqs}(\mathbf{c}), \\
\mathcal{J}(\mathbf{c}; W) &= \mathrm{Trim}_\gamma \left\{ \mathrm{mean}_{i \in \mathcal{I}_m} \, \mathrm{mean}_{q \in \mathcal{Q}_i^{\mathrm{all}}} \, \bar{\rho}_{iq}(\mathbf{c}; W) : m = 1, \ldots, M_{\mathrm{rec}} \right\}.
\end{aligned}
\tag{6}
$$

Each aggregation omits empty supports and non-finite scores. During axis scans, $\mathcal{J}(\mathbf{c}) = \mathcal{J}(\mathbf{c}; W(\mathbf{c}))$ uses each candidate's own valid pairs. For $a \in \{p, r\}$, we maximize $S_{b,a}^{\mathrm{cv}}(v) = \mathcal{J}(\mathbf{c}_b[a \leftarrow v])$ over $v \in \Omega_a$ and update that component.

Fixed-correspondence geometric yaw. Yaw instead uses normalized backbone grid features from t and its fixed ring neighbors $n \in \mathcal{R}_t$, without detection anchors. For each branch and pair, bidirectional back-projection over depths $\mathcal{D}$ and yaw candidates $\mathcal{H}_y$ retains locations on their own grids whenever a hypothesis projects validly into the other camera. These visibility unions restrict potential overlap; the hypotheses neither estimate depth nor establish matches.

Cosine mutual-nearest-neighbor matching with fixed similarity and ratio tests yields $\mathcal{M}_{i,b,t,n}$, held fixed across yaw candidates within the branch. Only the epipolar geometry induced by

$\mathbf{T}_{in}^{0}[\mathbf{T}_{it}(\mathbf{c}_b[y \leftarrow v])]^{-1}$ changes. Similarity-weighted Huber costs of square-root Sampson residuals are averaged over valid neighbors, observations, and sequences to obtain $E_{\text{geo}}^{(b)}(v)$ (Appendix A.4). We update

$$(\mathbf{c}_b)_y \leftarrow \left[\arg\min_{v \in \mathcal{H}_y} E_{\text{geo}}^{(b)}(v)\right] - \beta_t^{\text{clean}}, \tag{7}$$

where β_t^{clean} is the frozen camera-specific clean yaw offset.

Common-support state selection. Unlike candidate scans, final comparison uses the intersection of both states' valid pairs:

$$W_{iqs}^{\cap} = W_{iqs}(\mathbf{c}_R)W_{iqs}(\mathbf{c}_S), \qquad J_b = \mathcal{J}(\mathbf{c}_b; W^{\cap}), \quad b \in \{R, S\}. \tag{8}$$

This scores both states on identical visible support. We return $\mathbf{c}_S$ if $J_S > J_R$, otherwise $\mathbf{c}_R$, without modifying the live detector or its temporal memory.

4 Experiments

We evaluate the complete detector-evidence loop: unknown-camera diagnosis, rotational recovery, and restoration of the same frozen detector. Matched known-target baselines isolate recovery performance, while ablations examine the contribution of individual evidence sources and recovery constraints.

4.1 Protocol and Evaluation

Benchmark and access. We use a frozen ResNet-50 (He et al., 2016) StreamPETR (Wang et al., 2023) on nuScenes (Caesar et al., 2020). An observation and a sequence correspond to a nuScenes sample and scene, respectively. Each event applies one fixed rotational error to an unknown camera, while images and the remaining camera calibrations stay unchanged. Six cameras, eight roll–pitch–yaw sign patterns, and two per-axis magnitude ranges yield 96 conditions: 48 medium (M, 1–3°) and 48 large (L, 3–5°). The same 16 rotations are reused across splits.

The official nuScenes validation split contains 150 scenes, which we randomly partition once into two disjoint 75-scene subsets, Val-A and Val-B, using a fixed split seed. Val-A is used for development and ablation, whereas Val-B is reserved for post-development validation and downstream evaluation without further debugging or tuning. The official nuScenes test split is partitioned under a frozen metadata ordering: the first 75 scenes form Test-A for fully blind evaluation after the protocol is fixed, while the remainder is kept in reserve. Further partition and commissioning details are provided in Appendix B.1.

We set the Stage 1 diagnosis budget to $M_D = 40$ ordered sequences. Fast40 also uses $M_{\text{rec}} = 40$ for recovery, whereas HighAcc75 keeps the same diagnosis and extends recovery to $M_{\text{rec}} = 75$ by appending 35 sequences without re-encoding previous observations.

External baselines. All external baselines are given the faulty target camera and only perform rotational repair, whereas our method must identify it before recovery. They use its two fixed healthy ring neighbors, receive neither the perturbation nor ground-truth correction, and share the same first 40 ordered sequences as Fast40. XFeat (Potje et al., 2024) and DeDoDe-B (Edstedt et al., 2024a) share the same MNN matching, five-point essential estimation (Nistér, 2004), and $\mathrm{SO}(3)$ aggregation pipeline, differing only in the learned local feature front end. For FloSR (Kumar et al., 2024), we evaluate all released checkpoints on Val-A, select the one with the lowest rotational error, and freeze it for Val-B/Test-A. TESO (Moravec et al., 2026) follows its official online essential-matrix optimizer. Two-neighbor estimates are fused into a fixed target-camera correction. Full details are provided in Appendix B.2.

Fixed method settings. Unless otherwise specified, experiments use the default L3 configuration. The backbone level ℓ_B corresponds to stride 8, and anchors are read from decoder layer $\ell_A = 3$ (L3 under zero-based indexing) with $\mathcal{N} = 64$. Selected main and supplementary experiments additionally evaluate an L4 variant by setting $\ell_A = 4$ with all other settings unchanged. L4 results

are explicitly marked; unmarked results use L3. Robust aggregation uses $\gamma = 0.2$ and $\epsilon = 10^{-3}$. At deployment, Stage 1 uses $N_B = 4$ consecutive ten-sequence blocks and $\eta_{\text{pers}} = 2$. The threshold τ is commissioned from six clean development events. For each event j, G_{jc}^{clean} is obtained from its available ten-sequence blocks, and we define $z_j = \max_c G_{jc}^{\text{clean}}$ and freeze $\tau = \max_j z_j + \frac{1}{2}\,\text{IQR}(\{z_j\}) \approx 0.41305$. One commissioning event contains three rather than four blocks; all deployment diagnosis uses four blocks (Appendix A.2). Commissioning retains the six supported ring edges for the persistent graph.

For Stage 2, all axis scans use $\delta_a = 0.1^\circ$, with search radii $r_r = 12^\circ$ and $r_p = r_y = 8^\circ$. Sparse scoring uses $\lambda = 16$; NCC shifts span $\pm 6^\circ$ at 0.1° resolution, and $K_P = 30$ local yaw peaks are retained. Geometric yaw uses two fixed ring neighbors and depth hypotheses $\mathcal{D} = \{5, 10, 20, 40, 80\}\,\text{m}$. Mutual-nearest-neighbor matching uses a cosine-similarity threshold of 0.6, a ratio threshold of 0.95, and requires at least eight matches per neighbor. The frozen attention projections use $H = 8$ heads and $d_h = 32$. Stage 2 clean rank curves and camera-specific yaw offsets are frozen from the clean commissioning reference pool. All settings and reference statistics are fixed before Val-B and Test-A evaluation; further implementation details are provided in Appendix B.1.

Metrics. Stage 1 is evaluated by fault recall and camera Top-1 accuracy, where Top-1 is correct if the highest-scoring camera $\hat{c} = \arg\max_c G_c$ matches the true faulty camera. Rotational recovery is measured by $\text{SO}(3)$ geodesic error in degrees rather than Euler-component MAE. S@1 counts conditions with residual error at most 1°, while Positive counts conditions with a strict reduction from the initial error.

For Val-A and Val-B, each method first estimates one fixed correction from its recovery budget, which is then evaluated by rerunning the same frozen detector on all 75 sequences. To quantify downstream restoration, we report the performance recovery ratio (PRR), i.e., the fraction of corruption-induced performance loss recovered by the correction. For metric $\kappa \in \{\text{mAP}, \text{NDS}\}$,

$$\text{PRR}_\kappa(\mathcal{G}) = \frac{1}{|\mathcal{G}|} \sum_{j \in \mathcal{G}} \frac{\kappa_j^{\text{corr}} - \kappa_j^{\text{corrupt}}}{\kappa_j^{\text{clean}} - \kappa_j^{\text{corrupt}}}, \tag{9}$$

where j indexes perturbation conditions and $\mathcal{G}$ denotes the 48-condition M or L group, or all 96 conditions. PRR averages the recovery ratio computed separately for each condition rather than computing a ratio from condition-averaged metrics. Negative and above-100% values are retained. Test-A has no local detection labels, so downstream mAP/NDS and PRR are not reported.

4.2 From Fault Diagnosis to Rotational Recovery

Fault diagnosis and localization. Stage 1 achieves 100% fault recall and target-camera Top-1 accuracy across all 96 fault conditions in each split. No false alarms are observed on eight additional clean windows (Appendix B.3).

Matched rotational recovery. Under the matched 40-sequence budget, both Fast40 (L3) and Fast40 (L4) substantially outperform the known-target baselines (Table 1). Changing only the anchor readout from L3 to L4 further improves overall recovery: the mean/P90 residual decreases from $1.356^\circ/3.481^\circ$ to $0.957^\circ/1.637^\circ$ on Val-A, from $0.859^\circ/1.485^\circ$ to $0.692^\circ/1.327^\circ$ on Val-B, and from $1.223^\circ/2.361^\circ$ to $0.988^\circ/1.418^\circ$ on blind Test-A. The improvement is most pronounced for large faults, while Fast40 (L3) retains a slightly lower medium-fault mean on Val-A and Val-B. On Test-A, Fast40 (L4) reaches 78/96 S@1 conditions, compared with 48/96 for the strongest external baseline, DeDoDe-B. With the larger 75-sequence budget, HighAcc75 (L3) achieves the lowest Test-A mean/P90 residual of $0.894^\circ/1.284^\circ$.

4.3 Does Calibration Recovery Restore Perception?

Table 2 closes the loop back to the same frozen detector: each estimated correction is fixed and replayed over all 75 sequences without changing detector weights. Under the matched 40-sequence budget, Fast40 (L3) retains the strongest medium-fault restoration, recovering 96.11/96.21% of the mAP/NDS loss on Val-A and 97.85/96.37% on Val-B. Changing only the anchor readout from L3 to L4 substantially improves large-fault restoration, raising PRR from 78.59/79.73% to 91.21/91.20%

Table 1: **Rotational recovery by severity.** Errors are in degrees; S@1 and Pos. are group counts. †: known target. **Bold**: best displayed result per split; underline: best matched 40-sequence result (ties included).

		Medium (48)		Large (48)		All (96)		
Method	Seq.	Mean↓	S@1↑	Mean↓	S@1↑	Mean↓	P90↓	Pos.↑
Val-A								
XFeat+MNN+5pt E†	40	5.787	8	5.787	8	5.787	19.855	64
DeDoDe-B+MNN+5pt E†	40	3.371	24	3.371	24	3.371	11.070	72
FloSR-KITTI†	40	3.737	0	3.737	0	3.737	7.753	72
TESO†	40	3.270	0	6.998	0	5.134	7.858	65
Ours: Fast40 (L3)	40	**0.673**	40	2.039	31	1.356	3.481	91
Ours: Fast40 (L4)	40	0.792	**41**	**1.122**	**36**	**0.957**	**1.637**	93
Ours: HighAcc75 (L3)	75	0.839	38	1.519	32	1.179	1.817	**94**
Val-B								
XFeat+MNN+5pt E†	40	6.603	16	6.603	16	6.603	22.439	64
DeDoDe-B+MNN+5pt E†	40	2.691	24	2.691	24	2.691	8.146	73
FloSR-KITTI†	40	3.805	0	3.805	0	3.805	7.680	69
TESO†	40	3.114	0	6.988	0	5.051	7.641	70
Ours: Fast40 (L3)	40	**0.568**	**41**	1.149	34	0.859	1.485	94
Ours: Fast40 (L4)	40	0.576	**41**	**0.807**	**39**	**0.692**	**1.327**	**96**
Ours: HighAcc75 (L3)	75	0.615	38	1.100	33	0.858	1.548	94
Test-A								
XFeat+MNN+5pt E†	40	5.935	8	5.935	8	5.935	18.043	63
DeDoDe-B+MNN+5pt E†	40	3.361	24	3.361	24	3.361	10.798	72
FloSR-KITTI†	40	3.887	0	3.887	0	3.887	7.486	69
TESO†	40	3.006	0	6.964	0	4.985	7.545	77
Ours: Fast40 (L3)	40	0.782	35	1.665	29	1.223	2.361	93
Ours: Fast40 (L4)	40	0.570	42	1.405	**36**	0.988	1.418	**94**
Ours: HighAcc75 (L3)	75	**0.480**	**47**	**1.307**	34	**0.894**	**1.284**	93

on Val-A and from 92.16/91.60% to 95.81/94.84% on Val-B. HighAcc75 (L3) also improves large-fault restoration over Fast40 (L3), while Fast40 (L4) achieves the strongest restoration under the matched 40-sequence budget. Residual geometric error is strongly negatively associated with downstream recovery (Pearson $r = -0.959$ to -0.859; Appendix B.5), consistent with more accurate geometric recovery yielding stronger restoration of downstream perception.

4.4 What Makes the Detector Evidence Useful?

Task-aligned evidence and complementary constraints. Table 3 examines the main recovery mechanisms on Ablation24. Replacing confidence-selected anchors with random ones increases the median residual from $0.458°$ to $1.007°$ and reduces S@1 from 18/24 to 12/24. Removing cross-view pitch/roll refinement raises mean/P90 from $1.337°/3.013°$ to $3.604°/7.125°$ and reduces S@1 to 4/24, while geometric yaw provides complementary gains. Allowing the faulty camera to participate in source construction increases mean/P90 to $2.820°/6.870°$ and reduces S@1 to 8/24. Replacing sparse-assisted initialization and yaw proposals with zero initialization and a fixed yaw grid further degrades mean/P90 to $4.022°/9.233°$ and S@1 to 4/24.

Recovery-state and configuration diagnostics. Forcing the final correction to branch R or S gives mean residuals of $4.182°$ and $1.423°$, compared with $1.337°$ for the frozen dual-state common-support selector. Additional support, clean-reference, and decoder/stride analyses are provided in Appendix B.6; in particular, the statewise-support diagnostic is discussed separately from the frozen selector.

Portability and translation interference. Two additional StreamPETR checkpoints, including a different backbone, retain 11–12/12 correct diagnoses and localizations per split and substantial detection restoration (Table 20). With fixed approximately 3/5 cm per-axis translation interference,

Table 2: **Restoration of the same frozen detector.** mAP/NDS are reported $\times 100$ and PRR in percent; all corrections are evaluated by 75-sequence detector replay. L3/L4 denote the anchor-readout decoder layer; all other settings are unchanged. †: known target. **Bold**: best corrected result; underline: best matched 40-sequence result.

		Medium (48)				Large (48)			
Method	Seq.	mAP↑	NDS↑	PRR_{mAP} ↑	PRR_{NDS} ↑	mAP↑	NDS↑	PRR_{mAP} ↑	PRR_{NDS} ↑
Val-A									
Clean	–	41.79	52.45	–	–	41.79	52.45	–	–
Uncorrected	–	38.86	50.22	–	–	34.84	47.53	–	–
XFeat+MNN+5pt E†	40	37.79	49.86	-57.39	-33.73	37.79	49.86	44.38	49.62
DeDoDe-B+MNN+5pt E†	40	38.93	50.57	-9.64	4.01	38.93	50.57	60.68	63.54
FloSR-KITTI†	40	39.06	50.40	-47.52	-34.00	39.06	50.40	51.79	50.80
TESO†	40	39.23	50.48	13.64	11.07	35.10	47.68	4.09	3.25
Ours: Fast40 (L3)	40	**41.66**	**52.35**	**96.11**	**96.21**	40.27	51.39	78.59	79.73
Ours: Fast40 (L4)	40	41.37	52.16	90.11	90.48	**41.10**	**51.97**	**91.21**	**91.20**
Ours: HighAcc75 (L3)	75	41.52	52.24	92.87	93.08	40.83	51.75	87.34	86.87
Val-B									
Clean	–	45.04	55.15	–	–	45.04	55.15	–	–
Uncorrected	–	41.64	52.57	–	–	37.39	49.72	–	–
XFeat+MNN+5pt E†	40	40.85	52.35	-51.66	-30.14	40.85	52.35	44.16	48.37
DeDoDe-B+MNN+5pt E†	40	42.30	53.26	7.80	15.83	42.30	53.26	65.43	66.47
FloSR-KITTI†	40	41.85	52.70	-50.89	-37.04	41.85	52.70	47.84	46.21
TESO†	40	42.31	53.04	15.83	15.26	37.66	49.90	3.40	3.44
Ours: Fast40 (L3)	40	**44.99**	**55.07**	**97.85**	**96.37**	44.27	54.59	92.16	91.60
Ours: Fast40 (L4)	40	44.96	55.06	97.68	96.33	**44.65**	**54.82**	**95.81**	**94.84**
Ours: HighAcc75 (L3)	75	44.94	55.04	96.52	94.94	44.39	54.68	93.39	93.11

Table 3: **Core recovery ablations on Ablation24.** All settings use 40 sequences and known targets. Errors are in degrees; counts are out of 24. **Bold** marks the best result per column, including ties. The zero-init variant replaces sparse-assisted initialization with zero initialization and sparse yaw proposals with a fixed yaw grid.

	Mean↓		All (24)				
Variant	M (12)	L (12)	Mean↓	Med.↓	P90↓	S@1↑	Pos.↑
Full	**0.617**	2.057	**1.337**	**0.458**	3.013	**18**	23
Random anchors	0.801	1.993	1.397	1.007	**2.034**	12	**24**
w/o geometric yaw	1.275	**1.957**	1.616	0.845	4.339	13	**24**
w/o cross-view pitch/roll	2.521	4.686	3.604	3.012	7.125	4	21
Target-included source	1.325	4.315	2.820	1.700	6.870	8	21
Zero init. + fixed yaw grid	1.879	6.165	4.022	2.839	9.233	4	17

rotation-only recovery improves 44/48 conditions and recovers 84.32%/84.95% of mAP/NDS loss (Table 21). These tests concern checkpoint portability and rotational recovery, not full six-DoF calibration.

5 Conclusion

We formulate perception-driven extrinsic drift diagnosis and recovery and show that a frozen multi-camera 3D detector can serve not only as a consumer of camera calibration, but also as a source of evidence for maintaining it. TempCalib exploits detector-derived cross-view evidence to diagnose and localize a faulty camera, recover its rotational extrinsic, and restore downstream detection without retraining. Experiments on nuScenes demonstrate effective unknown-camera diagnosis and rotational recovery across validation and blind-test settings, with substantial restoration of the same frozen detector. These results establish detector-derived evidence as a practical signal for post-deployment geometric recovery.

AI USE STATEMENT

Generative AI tools were used as auxiliary aids during the preparation of this work, mainly for language polishing, code debugging, literature-related organization, and the refinement of some background descriptions and manuscript presentation. They were also occasionally used to assist with routine tasks such as LaTeX editing, figure preparation, and checking intermediate result summaries. In all such cases, the generated content served only as a working aid and was subsequently examined, revised, and verified by us before being incorporated into the manuscript or experimental workflow.

The research itself, including the formulation of the problem, development of the main ideas, methodological design, experimental planning, system implementation, and interpretation of the results, was carried out by us. Generative AI was not used to determine the central research direction, propose the core methodological contribution, generate experimental results, or replace scientific judgment in the design and evaluation of the method. All reported numbers originate from recorded experimental runs, and we independently checked the relevant code, results, technical claims, and citations used in the paper.

Generative AI thus remained limited to a supporting role in this work, primarily assisting with presentation, debugging, and the refinement of existing material. It did not independently contribute to the scientific novelty of the research. All AI-assisted content was reviewed and verified by us, and we take full responsibility for the final manuscript and for the correctness, rigor, and reproducibility of the reported work.

ETHICS STATEMENT

This study uses an existing public autonomous-driving benchmark and does not involve the collection of new data or the recruitment of human participants. All experiments are conducted offline on recorded sensor data, and we do not use the dataset to identify or analyze individuals.

Camera calibration directly affects geometric perception, and an incorrect correction may degrade downstream performance in safety-critical systems. Our work evaluates calibration diagnosis and correction proposals under controlled experimental settings and does not validate their automatic application on deployed vehicles. Any operational use would require additional safeguards before deployment. Any release of code or derived artifacts will follow the applicable dataset and third-party software licenses.

REPRODUCIBILITY STATEMENT

We provide the experimental settings required to reproduce the main results throughout the main paper and appendix, including dataset construction and splits, model and training configurations, hyperparameters, coordinate conventions, calibration perturbations, evidence extraction, search and matching procedures, recovery settings, and evaluation protocols. Additional implementation details and extended experimental results are included in the appendix to clarify the full experimental pipeline.

Upon acceptance, we plan to publicly release the source code, configuration files, evaluation scripts, and additional supporting materials to facilitate independent reproduction and verification of the reported results.

REFERENCES

Daniel Barath, Jana Noskova, Maksym Ivashechkin, and Jiri Matas. MAGSAC++, a fast, reliable and accurate robust estimator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1304–1312, 2020.

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

Dian Chen, Jie Li, Vitor Guizilini, Rares Andrei Ambrus, and Adrien Gaidon. Viewpoint equivariance for multi-view 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9213–9222, 2023.

Eric Dexheimer, Patrick Peluse, Jianhui Chen, James Pritts, and Michael Kaess. Information-theoretic online multi-camera extrinsic calibration. *IEEE Robotics and Automation Letters*, 7 (2):4757–4764, 2022.

Johan Edstedt, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. DeDoDe: Detect, don't describe—describe, don't detect for local feature matching. In *International Conference on 3D Vision*, 2024a.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19790–19800, 2024b.

Richard I. Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. BEVDet: High-performance multi-camera 3D object detection in bird-eye-view, 2021. arXiv:2112.11790.

Ganesh Iyer, R. Karnik Ram, J. Krishna Murthy, and K. Madhava Krishna. CalibNet: Geometrically supervised extrinsic calibration using 3D spatial transformer networks. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018.

Hongxiang Jiang, Wenming Meng, Hongmei Zhu, Qian Zhang, and Jihao Yin. Multi-camera calibration free BEV representation for 3D object detection, 2022. arXiv:2210.17252.

Takayuki Kanai, Igor Vasiljevic, Vitor Guizilini, Adrien Gaidon, and Rares Ambrus. Robust self-supervised extrinsic self-calibration. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1932–1939, 2023.

Anush Kumar, Fahim Mannan, Omid Hosseini Jafari, Shile Li, and Felix Heide. Flow-guided online stereo rectification for wide baseline stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15375–15385, 2024.

Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13906–13915, 2020.

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. BEVDepth: Acquisition of reliable depth for multi-view 3D object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1477–1485, 2023.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European Conference on Computer Vision*, 2022.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17627–17638, 2023.

Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: Position embedding transformation for multi-view 3D object detection. In *European Conference on Computer Vision*, 2022.

Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. PETRv2: A unified framework for 3D perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3262–3272, 2023.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

Zhaotong Luo, Guohang Yan, and Yikang Li. Calib-anything: Zero-training LiDAR-camera extrinsic calibration method using segment anything, 2023. arXiv:2306.02656.

Xudong Lv, Boya Wang, Ziwen Dou, Dong Ye, and Shuo Wang. LCCNet: LiDAR and camera self-calibration using cost volume network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2894–2901, 2021.

Jaroslav Moravec, Radim Šára, and Akihiro Sugimoto. TESO: Online tracking of essential matrix by stochastic optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28819–28827, 2026.

David Nistér. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):756–770, 2004. doi: 10.1109/TPAMI.2004.17.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023.

Bastian Pätzold, Simon Bultmann, and Sven Behnke. Online marker-free extrinsic camera calibration using person keypoint detections. In *DAGM German Conference on Pattern Recognition*, pp. 300–316, 2022.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In *European Conference on Computer Vision*, 2020.

Guilherme Potje, Felipe Cadar, André Araujo, Renato Martins, and Erickson R. Nascimento. XFeat: Accelerated features for lightweight image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2682–2691, 2024.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4938–4947, 2020.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8922–8931, 2021.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, pp. 9229–9248, 2020.

Ilir Tahiraj, Jeremialie Swadiryus, Felix Fent, and Markus Lienkamp. Cal or no cal? – real-time miscalibration detection of LiDAR and camera sensors, 2025. arXiv:2504.01040.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022a.

Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3621–3631, 2023.

Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. DETR3D: 3D object detection from multi-view images via 3D-to-2D queries. In *Proceedings of the 5th Conference on Robot Learning*, pp. 180–191, 2022b.

Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. RoboBEV: Towards robust bird's eye view perception under corruptions, 2023. arXiv:2304.06719.

Kaixin Xiong, Shi Gong, Xiaoqing Ye, Xiao Tan, Ji Wan, Errui Ding, Jingdong Wang, and Xiang Bai. CAPE: Camera view position embedding for multi-view 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21570–21579, 2023.

Weiduo Yuan, Jerry Li, Justin Yue, Divyank Shah, Konstantinos Karydis, and Hang Qiu. BEVCalib: LiDAR-camera calibration via geometry-guided bird's-eye view representation. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 4824–4836, 2025.

A SUPPLEMENTARY METHOD DETAILS

We follow the notation and settings in Section 4.1. Below, we provide additional details on the construction and aggregation of detector-derived calibration evidence.

A.1 EVIDENCE EXTRACTION AND COORDINATE CONVENTIONS

Decoder readout. For each 256×704 image, the ResNet-50 backbone supplies $\mathbf{F}_{ic}^{\ell_B} \in \mathbb{R}^{512\times32\times88}$ and the neck supplies $\mathbf{F}_{ic}^{N} \in \mathbb{R}^{256\times16\times44}$. Normal detection uses 644 base queries and 256 propagated queries. The Shared-Evidence Decoder and Target-Excluded Decoder each use only the base queries, without temporal memory. In each pass, the maximum sigmoid class probability at the final decoder layer selects the top-$\mathcal{N}$ query indices. Centers and hidden states at these indices are then read from layer ℓ_A; selection is not based on intermediate-layer confidence.

Projection and sampling. Following the nuScenes/StreamPETR geometry convention, 3D query centers are represented in the LiDAR coordinate frame, which serves only as the common reference frame for camera projection. This coordinate choice is inherited from the dataset representation and detector interface; our calibration method does not use LiDAR measurements, features, or point clouds. With homogeneous anchor $\bar{\mathbf{x}}_{iq} = [\mathbf{x}_{iq}^\top, 1]^\top$, projection gives $(\tilde{u}, \tilde{v}, z)^\top = \mathbf{P}_{ic}\bar{\mathbf{x}}_{iq}$, where

$$\mathbf{P}_{ic} = \mathbf{K}_{ic}\,[\mathbf{I}_3 \quad \mathbf{0}]\,\mathbf{T}_{ic}^0. \tag{10}$$

The corresponding backbone coordinates are

$$u_{icq}^{\ell_B} = \frac{\tilde{u}}{z}\frac{W_{\ell_B}}{W_{\text{img}}}, \qquad v_{icq}^{\ell_B} = \frac{\tilde{v}}{z}\frac{H_{\ell_B}}{H_{\text{img}}}. \tag{11}$$

These coordinates are used directly for continuous bilinear sampling; no half-cell offset is applied. This convention is distinct from the discrete grid-center coordinates used for geometric matching in Appendix A.4. Only finite, in-bounds projections with positive depth are considered valid. Descriptors are obtained by bilinear interpolation on the backbone feature map, with channel-wise ℓ_2 normalization applied before and after sampling. For candidate evaluation, only the target-camera transform is replaced by $\mathbf{T}_{it}(\mathbf{c})$. The rotation-only left action in equation 1 changes the extrinsic blocks to $\Delta\mathbf{R}\mathbf{R}^0$ and $\Delta\mathbf{R}\mathbf{t}^0$, while preserving the camera center in the reference frame, $-(\mathbf{R}^0)^\top\mathbf{t}^0$. Thus, candidate corrections introduce no independent camera-center displacement.

A.2 CLEAN REFERENCES AND GRAPH COMMISSIONING

Graph40 and Reference150 are train-derived development subsets containing 40 sequences/1,612 observations and 150 sequences/6,027 observations, respectively. They are mutually scene-disjoint and are each scene-disjoint from Val-A, Val-B, and Test-A. Their roles are summarized in Table 4. Historical Val-A examples contribute to threshold commissioning and graph-rule selection, consistent with its development role; Val-B and Test-A are not used for these choices.

Table 4: **Sources of fixed references.** All reference quantities and rule choices are frozen before post-development evaluation.

Quantity	Source and purpose
μ_e, σ_e	Graph40; normalize sequence-level edge consistency.
Stage 1 rule selection	Six clean events and 30 fault events, all from development-associated data.
τ	Six clean events: Graph40; Reference150 partitioned into $40+40+40+30$ sequences; one 40-sequence Val-A window. The 30-sequence Reference150 event uses three consecutive ten-sequence blocks.
$R_{\text{tar},a}^{\text{clean}}$, β_t^{clean}	Reference150; sparse curve alignment and camera-specific yaw-offset correction.

For each candidate edge $e \in \mathcal{E}_0$, we compute the observation-level consistency s_{ie} and aggregate it into the clean sequence-level score S_{me} as defined in the main text. An edge is retained in $\mathcal{E}$ if S_{me} is finite for at least two Graph40 clean sequences, the mean fraction of observations with valid

s_{ie} is at least 0.02, and the resulting reference statistics are finite. The clean mean μ_e and sample standard deviation σ_e are then computed from all valid S_{me} values for that edge, with ϵ used as the normalization floor. The retained edges form the ring Front–Front-Left–Back-Left–Back–Back-Right–Front-Right–Front. These adjacencies also define the healthy neighbors used for geometric yaw and the external baselines.

For $n > 0$ finite sorted scores $x_{(1)} \le \cdots \le x_{(n)}$, Trim_γ averages entries $x_{(k+1)}, \ldots, x_{(n-k)}$, where $k = \lfloor \gamma n \rfloor$. Each ten-sequence block therefore discards two scores from each tail when $\gamma = 0.2$. Deployment diagnosis always uses four such blocks. During development-time threshold commissioning, one 30-sequence clean event from Reference150 contains three consecutive ten-sequence blocks; its event score is the median of those three block scores, without padding or repetition. The six resulting clean-event maxima determine $\tau \approx 0.41305$ through the max-plus-half-IQR rule in Section 4.1.

A.3 Sparse proposal scoring and reference alignment

The Target-Excluded Decoder suppresses target features before region scoring, restricts selected regions to non-target cameras, and excludes target tokens from transformer memory. After localization, this pass is evaluated once per recovery observation. Final-layer confidence selects query indices for the layer-ℓ_A readout, as above. For query-position embedding $\mathbf{p}^Q_{iq}$, the fixed per-head query is

$$\mathbf{q}^{(-t)}_{iqh} = \mathbf{W}^Q_h(\mathbf{h}^{(-t)}_{iq} + \mathbf{p}^Q_{iq}), \tag{12}$$

where $\mathbf{W}^Q_h$ is taken from layer $\ell_A + 1$. The corresponding key projection acts on the target-neck feature sampled at the candidate projection, after the frozen positional encoding, spatial alignment, and gating operations. The query support is fixed under zero correction; candidate-invalid projections receive $-\lambda$ in equation 4 rather than being removed.

Temporal aggregation. Let $L_m = |\mathcal{I}_m|$ be the number of temporally ordered observations in sequence m. For $L_m > 6$, choose base positions $a_j = \mathrm{round}(j(L_m - 1)/5)$ for $j = 0, \ldots, 5$; otherwise use all positions. For each phase $\varphi \in \{0, 1, 2\}$, offset the base positions by φ, clip to the valid index range, and remove within-phase duplicates. Writing the resulting observation set as $\mathcal{I}^{\mathrm{sp}}_{m,\varphi}$,

$$C_{m,a}(v) = \mathrm{mean}_{\varphi \in \{0,1,2\}} \, \mathrm{mean}_{i \in \mathcal{I}^{\mathrm{sp}}_{m,\varphi}} \, S^{\mathrm{sp}}_i(\mathbf{0}[a \leftarrow v]). \tag{13}$$

Non-finite scores are omitted. Each sequence curve is converted to normalized ascending ordinal ranks over Ω_a, followed by equal-weight sequence averaging. Clean curves use the same sampling and aggregation.

Reference alignment. For each shift $\delta \in \Omega_{\mathrm{NCC}}$, let $\mathcal{O}_\delta$ denote the jointly finite overlap between $R_{\mathrm{tar},a}(v)$ and the shifted clean reference $R^{\mathrm{clean}}_{\mathrm{tar},a}(v - \delta)$. For $k \in \mathcal{O}_\delta$, define $x_k = R_{\mathrm{tar},a}(v_k)$ and $y_k(\delta) = R^{\mathrm{clean}}_{\mathrm{tar},a}(v_k - \delta)$. With overlap-specific means $\bar{x}$ and $\bar{y}_\delta$,

$$\mathrm{NCC}_a(\delta) = \frac{\sum_{k \in \mathcal{O}_\delta}(x_k - \bar{x})(y_k(\delta) - \bar{y}_\delta)}{\max\Big(\sqrt{\sum_{k \in \mathcal{O}_\delta}(x_k - \bar{x})^2 \sum_{k \in \mathcal{O}_\delta}(y_k(\delta) - \bar{y}_\delta)^2}, 10^{-12}\Big)}. \tag{14}$$

Shifts with fewer than five overlapping coordinates are discarded. The maximizing shift gives c^{sp}_a. For yaw, we additionally retain up to K_P strongest local maxima of the observed rank curve $R_{\mathrm{tar},y}$; together with zero and c^{sp}_y, these form $\mathcal{H}_y$.

A.4 Geometric yaw and final-state comparison

Fixed correspondences. For geometric yaw, feature locations are indexed by discrete backbone cells. A grid index (x_g, y_g) is associated with the corresponding cell-center image coordinate $((x_g + 0.5)W_{\mathrm{img}}/W_{\ell_B}, (y_g + 0.5)H_{\mathrm{img}}/H_{\ell_B})$. This discrete cell-center convention is used only for geometric yaw and is not treated as the inverse of the continuous sampling convention in equation 11. Given homogeneous pixel $\tilde{\mathbf{u}}_t$ and depth $d \in \mathcal{D}$, back-projection gives $\mathbf{X}_t(d) = d\mathbf{K}^{-1}_{it}\tilde{\mathbf{u}}_t$.

For branch b, neighbor $n \in \mathcal{R}_t$, and yaw candidate v, the relative transform is

$$\mathbf{T}^{(b)}_{i,n\leftarrow t}(v) = \mathbf{T}^{0}_{in}[\mathbf{T}_{it}(\mathbf{c}_b[y \leftarrow v])]^{-1} = \begin{bmatrix} \mathbf{R}^{(b)}_{i,nt}(v) & \mathbf{t}^{(b)}_{i,nt}(v) \\ \mathbf{0}^\top & 1 \end{bmatrix}. \tag{15}$$

A grid location is retained on its own image whenever any (v, d) projects validly into the other camera. Applying this in both directions defines the visibility unions used for feature matching. Matches are mutual nearest neighbors between normalized backbone descriptors, with best and second-best cosine similarities s_1, s_2 satisfying $s_1 \geq 0.6$ and $(1 - s_1)/\max(1 - s_2, 10^{-8}) \leq 0.95$. The absolute similarity threshold removes weak matches, while the best-to-second-best ratio suppresses ambiguous correspondences with similarly plausible alternatives. A neighbor contributes only with at least eight matches. Matches and similarity weights are fixed within each observation/branch/pair during yaw evaluation.

Epipolar energy. The candidate fundamental matrix is

$$\mathbf{F}^{\mathrm{epi}}_{i,b,n}(v) = \mathbf{K}_{in}^{-\top}[\mathbf{t}^{(b)}_{i,nt}(v)]_\times \mathbf{R}^{(b)}_{i,nt}(v)\mathbf{K}_{it}^{-1}, \tag{16}$$

where $[\mathbf{t}]_\times$ denotes the skew-symmetric cross-product matrix (Hartley & Zisserman, 2004). Let $\mathbf{F} = \mathbf{F}^{\mathrm{epi}}_{i,b,n}(v)$. For a fixed match j, the square-root Sampson residual is

$$d_j(\mathbf{F}) = \sqrt{\frac{(\tilde{\mathbf{u}}_{n,j}^\top \mathbf{F}\tilde{\mathbf{u}}_{t,j})^2}{\|(\mathbf{F}\tilde{\mathbf{u}}_{t,j})_{1:2}\|_2^2 + \|(\mathbf{F}^\top\tilde{\mathbf{u}}_{n,j})_{1:2}\|_2^2}}. \tag{17}$$

For $d \geq 0$, the Huber cost is $d^2/2$ when $d \leq \delta_H$ and $\delta_H(d - \delta_H/2)$ otherwise, with $\delta_H = 2$ pixels. Using $w_j = \max(s_j, 0)$ for match cosine similarity s_j, we obtain

$$\begin{aligned} E_{i,b,n}(v) &= \frac{\sum_{j\in\mathcal{M}_{i,b,t,n}} w_j \,\mathrm{Huber}_{\delta_H}(d_j(\mathbf{F}))}{\sum_{j\in\mathcal{M}_{i,b,t,n}} w_j}, \\ E^{(b)}_{\mathrm{geo}}(v) &= \mathrm{mean}_m \,\mathrm{mean}_{i\in\mathcal{I}_m}\,\mathrm{mean}_{n\in\mathcal{R}_t:\mathrm{valid}}\, E_{i,b,n}(v). \end{aligned} \tag{18}$$

This aggregation uses all valid recovery observations with equal sequence weights and no sequence-level trimming. Applying the same yaw search to Reference150 yields the camera-specific clean offset β_t^{clean}, which is subtracted from the selected raw yaw optimum in both branches, as in equation 7.

Branch states and support. Branches R and S start at zero and $\mathbf{c}^{\mathrm{sp}}$, respectively. Each pitch, yaw, or roll update retains the other two current components, so the initial sparse yaw and roll can also affect the first pitch update in branch S. Cross-view scans use candidate-dependent $W(\mathbf{c})$; only the final comparison uses $W^\cap$, the intersection of the two final states' valid anchor–source pairs. Both use the source–anchor–observation–sequence aggregation in equation 6. The selected correction is a proposal; it does not update network parameters or temporal memory.

B Additional Experimental Evidence

B.1 Evaluation protocol and metrics

Partitions and observation order. The 150 official nuScenes validation scenes are randomly partitioned once into two disjoint 75-scene subsets, Val-A and Val-B, using a fixed seed. Val-A supports development and ablation; Val-B is used for post-development validation and downstream evaluation without further tuning. Test-A contains the first 75 scenes in a frozen ordering of official-test metadata and is evaluated blind after protocol freezing. Table 5 gives the observation counts. Within each partition, observations are stably sorted by timestamp and scenes are ordered by first occurrence. Fast40 uses the first 40 scenes, grouped into four ten-scene blocks; HighAcc75 appends scenes 41–75. Each fixed correction is evaluated on all 75 scenes of Val-A or Val-B. Local Test-A detection labels are unavailable. Commissioning sources are specified once in Appendix A.2.

Table 5: **Evaluation partitions.** First-40 observations are shared by Fast40 and all matched external baselines.

Split	Role	Scenes	All obs.	First-40 obs.
Val-A	Development / ablation	75	3,017	1,605
Val-B	Post-development validation	75	3,002	1,592
Test-A	Blind test	75	2,989	1,586

Table 6: **Fixed rotational perturbations.** Angles and initial geodesic errors are in degrees; m/p denote negative/positive signs, ordered as roll, pitch, and yaw.

Level	Pattern	Roll	Pitch	Yaw	Initial error
Medium	mmm	-1.130	-1.015	-2.091	2.592
Medium	mmp	-2.675	-1.883	+1.034	3.417
Medium	mpm	-1.030	+1.809	-2.583	3.305
Medium	mpp	-1.382	+1.206	+2.908	3.450
Medium	pmm	+1.968	-2.659	-2.070	3.878
Medium	pmp	+1.145	-2.371	+2.495	3.643
Medium	ppm	+1.529	+2.612	-2.946	4.248
Medium	ppp	+2.876	+2.415	+2.167	4.305
Large	mmm	-3.631	-4.299	-4.886	7.540
Large	mmp	-4.808	-3.172	+3.728	6.787
Large	mpm	-3.289	+4.466	-3.313	6.393
Large	mpp	-4.836	+4.995	+4.992	8.678
Large	pmm	+3.662	-4.058	-4.557	7.031
Large	pmp	+4.656	-3.538	+3.328	6.798
Large	ppm	+3.682	+4.733	-4.931	7.858
Large	ppp	+3.839	+3.368	+4.804	6.932

Perturbations. Each fault event applies one fixed rotation to one camera. Synthetic drift uses $\mathbf{R}_z(y)\mathbf{R}_y(p)\mathbf{R}_x(r)$, whereas corrections follow equation 1. Each of the 16 triplets in Table 6 is applied to all six cameras on each split. The resulting 96 conditions reuse the split's images and are not independent scene collections. No independent translation is introduced in the main grid.

Additional metric definitions. For condition j, let $\mathbf{R}_j^\star$ denote the clean target rotation and $\mathbf{R}_j^s$ the target rotation under state $s \in \{\mathrm{corrupt}, \mathrm{corr}\}$. We measure the rotational residual by the $\mathrm{SO}(3)$ geodesic distance

$$e_j^s = \frac{180}{\pi} \arccos\left[\mathrm{clip}\left(\frac{\mathrm{tr}((\mathbf{R}_j^\star)^\top \mathbf{R}_j^s) - 1}{2}, -1, 1\right)\right]. \tag{19}$$

Thus, e_j^s is the minimum rotation angle between the corresponding target orientation and the clean orientation, rather than an Euler-component error. Mean, median, and P90 summarize these condition-level residuals, while Worst denotes the maximum. All conditions, including failed recoveries, remain in the denominator. S@0.5, S@1, and S@2 count residuals within the corresponding inclusive angular thresholds, while Positive requires $e_j^{\mathrm{corr}} < e_j^{\mathrm{corrupt}}$.

The aggregate geometric recovery is

$$\mathrm{Rec}_{\mathrm{geo}} = 100\left(1 - \frac{\sum_j e_j^{\mathrm{corr}}}{\sum_j e_j^{\mathrm{corrupt}}}\right). \tag{20}$$

PRR in equation 9 is computed condition-wise from paired full-precision clean, corrupted, and corrected outputs; conditions with a zero clean-to-corrupt gap are excluded, although none occurs in the reported main comparisons. Ratios are not clipped. For completeness, we also report the aggregate closure

$$\mathrm{Gap}_\kappa = 100\,\frac{\sum_j(\kappa_j^{\mathrm{corr}} - \kappa_j^{\mathrm{corrupt}})}{\sum_j(\kappa_j^{\mathrm{clean}} - \kappa_j^{\mathrm{corrupt}})}. \tag{21}$$

Unlike PRR, which gives each condition equal weight, this ratio of sums implicitly weights conditions by the magnitude of their original performance loss when those losses are positive.

B.2 External-baseline implementation

Common recovery protocol. All four external baselines are given the faulty target camera and use its two healthy ring neighbors. From the Fast40 observations in Table 5, each method estimates target–neighbor relative rotations, converts them to target-camera rotations using the healthy-neighbor extrinsics, and fuses the two estimates on $\mathrm{SO}(3)$. For estimated target rotation $\widehat{\mathbf{R}}_t$, the applied correction is $\Delta\mathbf{R}_t = \widehat{\mathbf{R}}_t(\mathbf{R}_t^0)^\top$, following equation 1. Only rotation is corrected; no independent translation is estimated. One valid neighbor is sufficient, while failure of both yields the identity correction without removing the condition. No method uses the clean target extrinsic, injected correction, TempCalib output, or downstream detection metric during per-condition estimation or fusion. Any development-time model selection is restricted to Val-A; the resulting configuration is fixed before Val-B and blind Test-A evaluation.

XFeat and DeDoDe-B. XFeat (Potje et al., 2024) uses its official pretrained sparse feature model. DeDoDe-B (Edstedt et al., 2024a) uses the original 3DV Detector-L and Descriptor-B weights rather than the later detector version. Neither is fine-tuned on nuScenes. Both retain up to 4,096 keypoints per image and use ℓ_2-normalized descriptors. To isolate the feature front end, we apply the same cosine mutual-nearest-neighbor matcher and geometric backend to both methods. Five-point essential estimation (Nistér, 2004) uses USAC-MAGSAC (Barath et al., 2020) with a one-pixel inlier threshold, followed by pose recovery. At least eight matches and eight cheirality inliers are required per observation. Valid target rotations for each camera pair are aggregated by an $\mathrm{SO}(3)$ geodesic medoid, and the two neighbor estimates are fused according to accumulated inlier support.

FloSR. FloSR (Kumar et al., 2024) is evaluated using its released pretrained models and official 6D rotation representation, without nuScenes fine-tuning. We compare the CARLA, Semi-Truck, and KITTI checkpoints on Val-A and select the one with the lowest mean geodesic residual; the resulting KITTI checkpoint is then fixed for Val-B and Test-A. FloSR is applied independently to the two target–neighbor pairs. Predicted relative rotations are converted to target-camera rotations, summarized within each pair by an $\mathrm{SO}(3)$ geodesic medoid, and fused equally across the two neighbors.

TESO. TESO (Moravec et al., 2026) follows its official online optimization on the essential manifold. Each frame uses 1,000 SIFT features (Lowe, 2004); normalized descriptors are matched bidirectionally with $K = 5$ nearest neighbors. We use $\sigma = 0.001$, update bound $\sigma/4$, $m_{\max} = 10$, and a ten-frame burn-in. The current target–neighbor pose initializes the essential matrix, after which the state is updated sequentially over Fast40 and only the final state is used. TESO is run independently for both ring neighbors, and their resulting rotations are fused according to processed-feature support.

B.3 Fault diagnosis and clean checks

Table 7 reports the final diagnosis results. Camera ranks are computed among all six cameras; all Full96 faults are detected, so conditional and unconditional localization counts coincide. We additionally evaluate 18 previously unseen perturbation conditions on the same Val-A scene pool, achieving 18/18 fault detection and target-camera localization. This provides a condition-level, rather than scene-level, holdout check.

Table 7: **Full96 diagnosis and localization.** All splits use M_D sequences; margin is the minimum event score minus τ.

Split	Fault recall↑	Top-1↑	Top-2↑	MRR↑	Min. margin↑
Val-A	96/96	96/96	96/96	1.000	0.527
Val-B	96/96	96/96	96/96	1.000	0.811
Test-A	96/96	96/96	96/96	1.000	0.841

The six threshold-commissioning events are not treated as independent false-alarm tests. On eight additional overlapping 40-sequence clean windows, the frozen detector produces no false alarms.

B.4 Additional rotational-recovery results

Tables 8 and 9 extend the main recovery results with severity-specific statistics and additional success thresholds for Fast40 (L3), Fast40 (L4), and HighAcc75 (L3). All configurations are evaluated on the same 96-condition grids, and unsuccessful corrections are retained in the reported aggregates.

Under the matched 40-sequence budget, changing the anchor readout from L3 to L4 consistently improves aggregate recovery across all three splits. In particular, S@1 increases from 71/96 to 77/96 on Val-A, from 75/96 to 80/96 on Val-B, and from 64/96 to 78/96 on Test-A, while geometric recovery increases from 75.03% to 82.37%, from 84.18% to 87.26%, and from 77.46% to 81.81%, respectively. The severity-specific results further show that these gains are most pronounced for large faults, consistent with the main comparison.

Table 8: **Severity-specific TempCalib recovery.** Errors are degrees; each M/L row contains 48 conditions.

Split	Method	Level	Mean↓	Med.↓	P90↓	Worst↓	S@1↑	Pos.↑
Val-A	Fast40 (L3)	M	0.673	0.479	1.378	3.302	40/48	48/48
Val-A	Fast40 (L3)	L	2.039	0.634	7.141	10.345	31/48	43/48
Val-A	Fast40 (L4)	M	0.792	0.446	1.223	7.427	41/48	46/48
Val-A	Fast40 (L4)	L	1.122	0.525	2.622	8.304	36/48	47/48
Val-A	HighAcc75 (L3)	M	0.839	0.512	1.600	9.050	38/48	47/48
Val-A	HighAcc75 (L3)	L	1.519	0.728	5.068	9.682	32/48	47/48
Val-B	Fast40 (L3)	M	0.568	0.432	1.185	2.068	41/48	48/48
Val-B	Fast40 (L3)	L	1.149	0.623	1.979	7.996	34/48	46/48
Val-B	Fast40 (L4)	M	0.576	0.458	1.307	1.629	41/48	48/48
Val-B	Fast40 (L4)	L	0.807	0.517	1.344	5.022	39/48	48/48
Val-B	HighAcc75 (L3)	M	0.615	0.442	1.432	1.997	38/48	48/48
Val-B	HighAcc75 (L3)	L	1.100	0.627	2.113	8.085	33/48	46/48
Test-A	Fast40 (L3)	M	0.782	0.556	1.785	2.945	35/48	48/48
Test-A	Fast40 (L3)	L	1.665	0.618	2.955	13.603	29/48	45/48
Test-A	Fast40 (L4)	M	0.570	0.489	1.161	1.503	42/48	48/48
Test-A	Fast40 (L4)	L	1.405	0.627	2.565	14.238	36/48	46/48
Test-A	HighAcc75 (L3)	M	0.480	0.437	0.774	2.953	47/48	48/48
Test-A	HighAcc75 (L3)	L	1.307	0.572	2.901	10.461	34/48	45/48

Table 9: **Additional success thresholds.** Geometric recovery follows equation 20.

Split	Method	S@0.5↑	S@1↑	S@2↑	Geo rec. (%)↑
Val-A	Fast40 (L3)	45/96	71/96	81/96	75.03
Val-A	Fast40 (L4)	45/96	77/96	88/96	82.37
Val-A	HighAcc75 (L3)	40/96	70/96	87/96	78.29
Val-B	Fast40 (L3)	49/96	75/96	90/96	84.18
Val-B	Fast40 (L4)	50/96	80/96	93/96	87.26
Val-B	HighAcc75 (L3)	46/96	71/96	90/96	84.20
Test-A	Fast40 (L3)	38/96	64/96	81/96	77.46
Test-A	Fast40 (L4)	40/96	78/96	89/96	81.81
Test-A	HighAcc75 (L3)	46/96	81/96	89/96	83.54

The effect of a larger recovery budget is evaluated separately within the L3 configuration. Moving from Fast40 (L3) to HighAcc75 (L3) improves/degrades/leaves unchanged 45/42/9 conditions on Val-A, 42/44/10 on Val-B, and 58/33/5 on Test-A. Thus, additional observations improve the aggregate Test-A mean and tail error but do not benefit every condition. Figure 3 shows the full Test-A HighAcc75 (L3) grid, including three non-positive corrections despite its stronger aggregate performance.

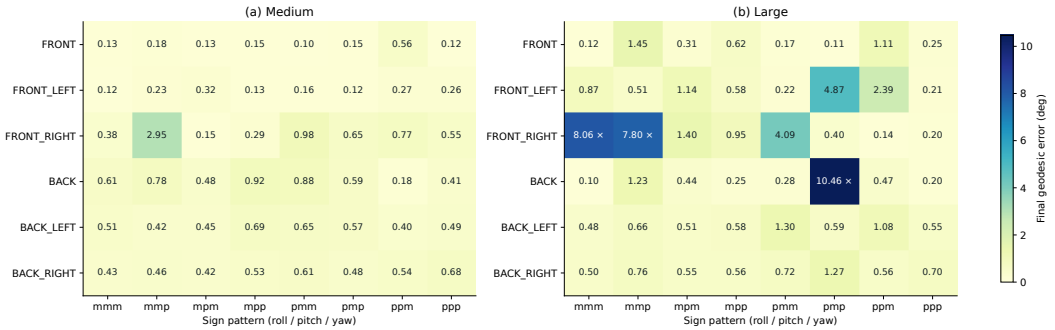


Figure 3: **All 96 Test-A HighAcc75 residuals.** Rows are cameras and columns are perturbations. Values are geodesic degrees; × marks residuals not smaller than their initial error.

B.5 Downstream restoration and geometry

Tables 10 and 11 provide additional downstream analyses beyond the mAP/NDS results in Table 2. Table 10 reports the standard nuScenes true-positive errors: mATE, mASE, mAOE, mAVE, and mAAE, which measure translation, scale, orientation, velocity, and attribute errors, respectively (Caesar et al., 2020); lower values are better. Across both Val-A and Val-B, all three TempCalib configurations substantially reduce the degradation caused by rotational corruption and bring these errors back toward their clean values.

Table 10: **True-positive detection errors for TempCalib.** Each row averages 96 full-split evaluations. Clean/corrupted values are shared by the three methods within each split and are shown once.

Method	State	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
Val-A						
Clean	Clean	0.6190	0.2749	0.4996	0.2746	0.1761
Corrupted	Corrupted	0.6950	0.2800	0.5148	0.2848	0.1805
Fast40 (L3)	Corrected	0.6295	0.2758	0.5021	0.2772	0.1770
Fast40 (L4)	Corrected	0.6273	0.2751	0.5002	0.2757	0.1768
HighAcc75 (L3)	Corrected	0.6278	0.2756	0.5023	0.2764	0.1769
Val-B						
Clean	Clean	0.5893	0.2675	0.4013	0.2810	0.1982
Corrupted	Corrupted	0.6704	0.2699	0.4249	0.2932	0.2032
Fast40 (L3)	Corrected	0.5963	0.2673	0.4054	0.2805	0.1987
Fast40 (L4)	Corrected	0.5951	0.2674	0.4046	0.2801	0.1990
HighAcc75 (L3)	Corrected	0.5954	0.2674	0.4054	0.2805	0.1986

Table 11: **All-condition downstream aggregates for TempCalib.** mAP/NDS use native $[0, 1]$ units; PRR and Gap are percentages. Gap follows equation 21.

Split	Method	mAP↑	NDS↑	PRR_{mAP} ↑	Gap_{mAP} ↑	PRR_{NDS} ↑	Gap_{NDS} ↑
Val-A	Fast40 (L3)	0.409673	0.518684	87.35	83.41	87.97	83.70
Val-A	Fast40 (L4)	0.412355	0.520664	90.66	88.84	90.84	89.23
Val-A	HighAcc75 (L3)	0.411754	0.519987	90.11	87.62	89.98	87.34
Val-B	Fast40 (L3)	0.446276	0.548309	95.00	92.52	93.98	92.09
Val-B	Fast40 (L4)	0.448052	0.549415	96.74	95.74	95.58	94.84
Val-B	HighAcc75 (L3)	0.446665	0.548608	94.95	93.23	94.02	92.83

Table 11 aggregates downstream restoration over all 96 conditions in each split. Fast40 (L4) achieves the strongest overall restoration on both Val-A and Val-B, reaching 90.66/90.84% macro PRR for mAP/NDS on Val-A and 96.74/95.58% on Val-B. The same trend is reflected by the corresponding Gap metrics. Together with the geometric recovery results, these measurements show that the im-

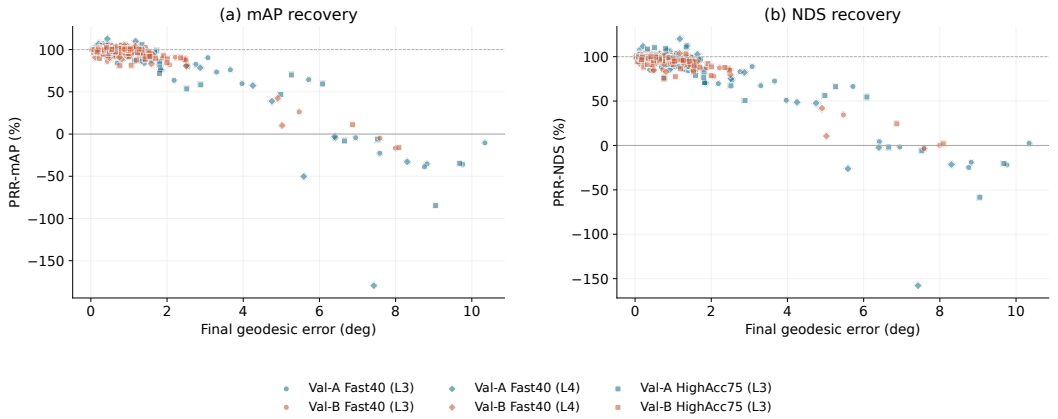


Figure 4: **Rotational error and perception recovery.** The analysis includes six split/method groups: Val-A and Val-B, each with Fast40 (L3), Fast40 (L4), and HighAcc75 (L3). Each group contains 96 conditions. Negative and above-100% PRR values are retained.

proved L4 readout translates consistently into stronger restoration of the frozen detector rather than only lower rotation residuals.

Across the six split/method groups, final rotational residual is negatively associated with condition-level mAP/NDS recovery. Pearson correlations range from -0.959 to -0.859, and Spearman correlations from -0.721 to -0.305 (Figure 4). These associations are consistent with the trend that more accurate geometric recovery yields stronger downstream restoration.

B.6 Ablation protocols and reference diagnostics

Component interventions. Ablation24 is a balanced Val-A subset containing 24 conditions: for each of the six cameras, we select two medium and two large rotational faults. All variants use the same 40-sequence/1,605-observation budget and are evaluated with known targets to isolate the recovery mechanisms.

Random-anchor ablation replaces confidence-ranked queries with $\mathcal{N}$ random queries while keeping the remaining pipeline unchanged. Removing cross-view pitch/roll skips only those refinement steps, whereas removing geometric yaw skips only the yaw refinement; all other branch operations and final selection are retained. To evaluate target exclusion, the target-included variant allows the faulty camera to participate in source construction and rebuilds the corresponding clean reference from the same Reference150 pool. To evaluate sparse-assisted proposal generation, the zero-initialization variant removes the sparse proposal and replaces its yaw candidates with a fixed grid over $[-8^\circ, 8^\circ]$ at 0.5° intervals, while retaining the subsequent pitch–yaw–roll refinement. Since this intervention changes both initialization and yaw candidate generation, it is treated as a mechanism-level replacement rather than an isolated sparse-proposal removal.

The paired comparisons in Table 3 show that the full method improves 17/24 conditions over random anchors, 17/24 over removing geometric yaw, 20/24 over removing cross-view pitch/roll, 18/24 over the target-included source, and 20/24 over the zero-initialized fixed-grid variant. The aggregate results therefore support the combined roles of confidence-based anchor selection, complementary cross-view and geometric refinement, target-excluded source construction, and sparse-assisted proposal generation.

Recovery-state and support diagnostics. We further isolate the contribution of the two recovery trajectories and their final comparison rule on the same Ablation24 conditions. The frozen solver produces two final states, R and S, which are normally compared using their common visible support. R-only and S-only commit one of these precomputed states directly, without changing candidate generation or refinement. We additionally evaluate a post-hoc statewise-support diagnostic, in which each state is scored on its own valid support rather than their intersection.

Table 12: **Recovery-state and final-support diagnostics on Ablation24.** All variants reuse the same Fast40 branch states and differ only in the committed state or final support used for comparison. Errors are in degrees; counts are out of 24. ‡ denotes a post-hoc diagnostic and is not part of the frozen TempCalib configuration.

Variant	M Mean↓	L Mean↓	Mean↓	Med.↓	P90↓	S@1↑	Pos.↑
Dual / common support (frozen)	0.617	2.057	1.337	0.458	3.013	18/24	23/24
R-only	1.972	6.391	4.182	3.026	9.602	3/24	18/24
S-only	1.378	1.468	1.423	0.380	4.683	17/24	22/24
Dual / statewise support‡	0.617	1.393	1.005	0.458	1.712	18/24	24/24

R-only is substantially weaker, while S-only approaches the frozen dual-state result, indicating that the sparse-assisted trajectory contributes most of the high-precision solutions but does not subsume the refinement-only branch. The statewise-support diagnostic agrees with common-support selection in 23/24 conditions. The sole disagreement occurs on one large-fault case, where the alternative branch reduces the residual from 9.760° to 1.803°. This single case accounts for most of the aggregate reduction from 1.337° to 1.005°, showing that the final selection can be sensitive to the support definition.

Disjoint clean references. To assess sensitivity to the Stage 2 clean reference pool, we split Reference150 into two deterministic, scene-disjoint subsets, Ref-A and Ref-B, containing 75 sequences and 3,009/3,018 observations, respectively. Only $R^{\text{clean}}_{\text{tar},a}$ and β^{clean}_t are recomputed; all other settings remain fixed. As shown in Table 13, replacing the full reference changes the overall mean residual from 1.337° to 1.377°/1.428°, while both subsets retain 23/24 positive recoveries. Ref-A and Ref-B select the same final branch in 20/24 conditions, with a mean absolute residual difference of 0.305°, indicating limited sensitivity to the particular clean-reference subset.

Table 13: **Clean-reference sensitivity on Val-A Ablation24.** Ref-A and Ref-B are scene-disjoint halves of Reference150. Only the Stage 2 clean rank references and camera-specific yaw offsets are recomputed; all other settings are fixed. Errors are in degrees; counts are out of 24.

Reference	M Mean↓	L Mean↓	Mean↓	Med.↓	P90↓	S@1↑	Pos.↑
Reference150	0.617	2.057	1.337	0.458	3.013	18/24	23/24
Ref-A	0.607	2.147	1.377	0.612	3.295	18/24	23/24
Ref-B	0.583	2.272	1.428	0.542	3.020	17/24	23/24

Clean-reference alignment. The sparse-response curves exhibit strong alignment with their clean references after shift compensation. As shown in Table 14, the median best-shift NCC exceeds 0.93 across all axes and severities, with roll showing the strongest agreement. Alignment remains high overall under large faults, although pitch shows greater lower-tail variation and a few pitch/yaw cases reach the NCC search boundary. These results support the use of clean-reference alignment for sparse proposal construction while also reflecting axis-dependent variation in the response curves.

Table 14: **Clean-reference alignment diagnostic on Ablation24.** Best-shift NCC is summarized over 12 conditions per severity and axis. Boundary hits count optima at the NCC search limit.

Axis	M Med.↑	M P10↑	L Med.↑	L P10↑	Boundary M/L↓
Pitch	0.950	0.788	0.931	0.704	0/12 / 1/12
Yaw	0.965	0.848	0.981	0.880	0/12 / 2/12
Roll	0.988	0.956	0.991	0.923	0/12 / 0/12

Camera-specific yaw-offset diagnostic. We further isolate the camera-specific clean yaw offset while keeping the Full candidate set and selected branch fixed. Table 15 reports yaw-axis errors before and after subtracting the frozen clean offset. The offset reduces the overall yaw MAE from 0.828° to 0.719°, while S@1 remains unchanged at 19/24. Its effect is less consistent in the tail,

with overall P90 increasing from 2.194° to 2.483°. Overall, the clean offset improves average yaw accuracy on this subset, while its benefit varies across conditions.

Table 15: **Camera-specific yaw-offset diagnostic on Ablation24.** Candidates and final branch selection are fixed; errors are yaw-axis absolute errors rather than final $\mathrm{SO}(3)$ residuals.

Level	N	Yaw MAE↓		P90↓		Yaw S@1 (w/o / Full)
		w/o offset	Full	w/o offset	Full	
Medium	12	0.686	0.481	1.169	1.010	10/12 / 10/12
Large	12	0.969	0.957	2.371	3.027	9/12 / 9/12
All	24	0.828	0.719	2.194	2.483	19/24 / 19/24

Decoder readout and feature stride. Table 16 varies the decoder readout at fixed stride 8 and the image-feature stride at fixed L3, using the same Ablation24 conditions, known targets, and Fast40 budget. Each representation uses its corresponding clean reference constructed from the same commissioning set. Among the matched L1–L4 readouts, L4 gives the lowest overall mean and P90 (0.921° and 1.278°) and the highest S@1 (20/24), whereas L3 has the lowest median (0.458°) and medium-fault mean (0.617°). L5 yields a lower overall mean, but has no subsequent decoder layer and reuses the final native query projection; it is therefore reported as a terminal-state diagnostic rather than a matched readout replacement. With L3 fixed, stride 8 achieves the lowest median and highest S@1, whereas stride 16 reduces mean/P90 to $1.199^\circ/2.046^\circ$ and achieves 24/24 positive recoveries. The stride study thus characterizes the L3 configuration, not an optimized stride choice for L4.

B.7 Recovery-budget sensitivity and selection diagnostics

Table 17 varies only the recovery budget on Val-A, while the diagnosis budget remains M_D. The 20-sequence setting therefore changes the recovery stage only and does not constitute an automatic 20-sequence system. Recovery quality is not monotonic in the observation budget. Compared with Fast40, the 20-sequence setting has lower mean and P90 residuals but a larger worst-case error and fewer S@1 successes. HighAcc75 further reduces the mean, P90, and worst-case residual, although its S@1 count remains comparable to Fast40.

Oracle selection. To separate errors from final branch selection and earlier candidate selection, we consider two diagnostic oracles. Oracle R/S selects the lower-error final branch using the clean target rotation. Oracle Candidate instead uses the clean rotation to select among the candidates generated at each refinement stage, while preserving the frozen candidate sets, branch initializations, and pitch–yaw–roll order. Neither oracle enlarges the search space or combines axis-wise optima from different trajectories, and both are used only to measure selection headroom.

Table 18 shows that Oracle R/S reduces the residual by choosing between the two final branches, whereas Oracle Candidate reveals substantially larger headroom in the earlier stage-wise selections.

A separate branch-level check on the three non-positive Test-A HighAcc75 conditions finds that two contain a positive alternative final branch, whereas neither branch is positive in the remaining case. This distinguishes failures caused by final-state selection from cases in which the generated branch outputs themselves are insufficient.

B.8 Rotation and translation sensitivity

We evaluate the pre-recovery sensitivity of the frozen detector along a fixed oblique direction $[0.547537, 0.480360, 0.685170]$. For rotation, the listed magnitudes specify the angular perturbation about this direction; for translation, they specify the displacement magnitude along the same unit direction. Each magnitude is applied to all six cameras with both signs, yielding 12 full-split evaluations. Table 19 reports the results referenced in the Introduction. This experiment characterizes detector sensitivity before recovery; the rotational and translational magnitudes are not intended to represent matched physical severity.

Table 16: **Decoder readout and image-feature stride ablations.** All variants use the same Val-A Ablation24 conditions (12 medium and 12 large), Fast40, known targets, and the frozen detector. Errors are in degrees; S@1 and Pos. are counts out of 24. Each derived representation uses its own clean reference from the same commissioning set. ‡L5 has no subsequent decoder layer and reuses the final native query projection; it is a terminal-state boundary diagnostic, not a strictly matched L1–L4 replacement.

Variant	M Mean↓	L Mean↓	All Mean↓	Med.↓	P90↓	Worst↓	S@1↑	Pos.↑
Decoder readout; image stride-8 fixed								
L1	0.870	3.288	2.079	0.797	6.191	10.362	13/24	22/24
L2	1.036	1.673	1.354	0.725	3.315	7.643	17/24	22/24
L3 (frozen)	0.617	2.057	1.337	0.458	3.013	9.760	18/24	23/24
L4	0.910	0.931	0.921	0.583	1.278	5.588	20/24	23/24
L5‡	0.562	1.024	0.793	0.533	1.276	5.878	19/24	24/24
Image-feature stride; decoder L3 fixed								
Stride-4	0.616	2.315	1.465	0.499	4.171	8.450	17/24	22/24
Stride-8 (frozen)	0.617	2.057	1.337	0.458	3.013	9.760	18/24	23/24
Stride-16	0.724	1.675	1.199	0.671	2.046	7.484	17/24	24/24

Table 17: **Recovery-budget ablation on Val-A.** Errors are in degrees; diagnosis always uses M_D sequences.

Recovery Seq.	Mean↓	Med.↓	P90↓	Worst↓	S@1↑	Pos.↑
20	1.265	0.522	2.764	13.866	67/96	93/96
40	1.356	0.511	3.481	10.345	71/96	91/96
75	1.179	0.604	1.817	9.682	70/96	94/96

Table 18: **Ground-truth selection diagnostics on Full96.** Mean, median, P90, and Worst are in degrees.

Selection	Mean↓	Med.↓	P90↓	Worst↓	Pos.↑
Val-A					
Fast40 (L3)	1.356	0.511	3.481	10.345	91/96
Oracle R/S	0.939	0.474	1.888	6.950	96/96
Oracle Candidate	0.215	0.138	0.357	2.228	96/96
Val-B					
Fast40 (L3)	0.859	0.483	1.485	7.996	94/96
Oracle R/S	0.790	0.468	1.436	7.587	95/96
Oracle Candidate	0.176	0.151	0.312	0.672	96/96
Test-A					
Fast40 (L3)	1.223	0.583	2.361	13.603	93/96
Oracle R/S	0.883	0.577	1.957	6.133	96/96
Oracle Candidate	0.176	0.110	0.353	1.504	96/96

Table 19: **Detection sensitivity before recovery.** Deltas are perturbed minus clean in native $[0, 1]$ units; each row averages 12 conditions.

Perturbation	Magnitude	mAP	ΔmAP	NDS	ΔNDS
Rotation	1°	0.415657	-0.002206	0.522543	-0.001973
Rotation	2°	0.408356	-0.009508	0.516887	-0.007630
Rotation	5°	0.377076	-0.040788	0.494231	-0.030285
Translation	2 cm	0.417930	+0.000066	0.524578	+0.000062
Translation	5 cm	0.417883	+0.000019	0.524541	+0.000024
Translation	10 cm	0.417910	+0.000046	0.524509	-0.000008

B.9 PORTABILITY ACROSS CHECKPOINTS AND BACKBONES

To assess portability beyond the default detector checkpoint, we evaluate two additional StreamPETR configurations: R50–428q–60e and a VoVNetV2-99-eSE–24e variant (Lee & Park, 2020). Each configuration is tested on 12 conditions per split, comprising one medium and one large fault for each camera. As shown in Table 20, Stage 1 detects and localizes 11/12 conditions for R50 on Val-A and 12/12 for all remaining checkpoint/split combinations. For the single missed R50 Val-A condition, rotational recovery is evaluated separately with the target camera provided; this case contributes to the recovery metrics but not to the automatic diagnosis/localization count. All reported recovery trials reduce rotational error. Full Val-A detector replay further shows substantial downstream restoration for both configurations, supporting portability across StreamPETR checkpoints and backbones.

Table 20: **Portability across StreamPETR checkpoints and backbones.** Errors are in degrees. Val-A downstream scores use full 75-sequence replay; mAP/NDS are reported $\times 100$ and PRR in percent.

(a) Diagnosis and recovery

Split	Checkpoint	Det.↑	Top-1↑	Mean↓	Med.↓	P90↓	Worst↓	S@1↑	Pos.↑
Val-A	R50–428q–60e	11/12	11/12	1.202	0.726	1.858	5.992	8/12	12/12
Val-A	VoV-99-eSE–24e	12/12	12/12	2.017	1.002	5.638	6.772	6/12	12/12
Val-B	R50–428q–60e	12/12	12/12	1.015	0.595	1.410	5.866	10/12	12/12
Val-B	VoV-99-eSE–24e	12/12	12/12	1.411	0.945	2.099	5.952	6/12	12/12

(b) Val-A detection restoration

	Clean		Corrupted		Corrected		Macro PRR	
Checkpoint	mAP↑	NDS↑	mAP↑	NDS↑	mAP↑	NDS↑	mAP↑	NDS↑
R50–428q–60e	43.72	53.81	39.53	50.88	42.95	53.25	89.45	87.04
VoV-99-eSE–24e	46.15	55.57	43.42	53.70	44.96	54.78	79.38	104.76

Det. and Top-1 denote fault detection and target-camera localization. The R50 Val-A recovery metrics include one recovery-only case evaluated with the target camera provided after unsuccessful automatic localization.

B.10 ROTATIONAL RECOVERY WITH TRANSLATION INTERFERENCE

The 3 cm and 5 cm settings denote approximately 3 cm and 5 cm *per-axis* translation offsets in the faulty target-camera frame, rather than the norm of the translation vector. The injected vectors are $\Delta\mathbf{t} = (3.0, -2.7, 3.3)^\top$ cm and $(5.0, -4.5, 5.5)^\top$ cm, with norms of approximately 5.21 cm and 8.69 cm, respectively. Translation is applied after the rotational perturbation:

$$\mathbf{T}_{\text{mixed}} = \begin{bmatrix} \mathbf{I}_3 & \Delta\mathbf{t} \\ \mathbf{0}^\top & 1 \end{bmatrix} \mathbf{T}_{\text{rot}}, \tag{22}$$

where $\mathbf{T}_{\text{rot}}$ denotes the rotation-corrupted LiDAR-to-camera transform. The left action introduces a translation offset in the target-camera frame while leaving the rotation block unchanged. We evaluate 24 rotational conditions drawn from Ablation24, covering both medium and large faults across all six cameras. Each condition is paired with both translation levels, so the ≈ 3 cm/axis and ≈ 5 cm/axis groups share the same rotational perturbations and differ only in translation magnitude. The perturbation remains fixed within each event, with translation expressed in the local frame of the faulty camera. TempCalib remains rotation-only and does not estimate an independent translation correction.

Table 21 compares each mixed perturbation with its paired rotation-only condition. Across all 48 conditions, mean rotational residual increases from 1.337° to 1.532°, while 44/48 conditions still improve and 36/48 reach S@1. Full75 replay recovers 84.32%/84.95% of the mAP/NDS loss. These results demonstrate robustness of rotational recovery under the tested unmodeled translation interference, rather than translation recovery or full six-DoF calibration.

Table 21: **Rotational recovery under per-axis translation interference.** The first two rows group the 48 conditions by translation level, while the next two regroup the same conditions by rotational severity. Rotational errors are in degrees; mAP/NDS are reported $\times 100$ and PRR in percent.

(a) Rotational recovery

Group	N	Rot.-only Mean↓	Mixed Mean↓	Med.↓	P90↓	Worst↓	Δ_{trans} ↓	S@1↑	Pos.↑
≈ 3 cm/axis	24	1.337	1.580	0.585	4.535	10.750	+0.244	18/24	22/24
≈ 5 cm/axis	24	1.337	1.484	0.682	3.612	10.331	+0.147	18/24	22/24
Medium	24	0.617	1.030	0.619	1.444	5.588	+0.413	20/24	22/24
Large	24	2.057	2.034	0.761	6.433	10.750	-0.022	16/24	22/24
All	48	1.337	1.532	0.663	4.615	10.750	+0.195	36/48	44/48

(b) Full75 detection restoration

		Corrupted		Corrected		Macro PRR	
Group	N	mAP↑	NDS↑	mAP↑	NDS↑	mAP↑	NDS↑
≈ 3 cm/axis	24	36.73	48.69	40.49	51.53	82.97	84.01
≈ 5 cm/axis	24	36.75	48.69	40.68	51.64	85.66	85.90
Medium	24	38.63	50.05	41.14	52.01	89.07	90.40
Large	24	34.85	47.33	40.03	51.16	79.56	79.51
All	48	36.74	48.69	40.59	51.58	84.32	84.95

Clean mAP/NDS are 41.79/52.45. Rot.-only denotes the residual for the paired rotational condition without translation interference, and Δ_{trans} is the mixed-condition residual minus this rotation-only residual; positive values indicate degradation. All 48 compound conditions are retained in the evaluation.