FROM EXPLANATION TO CORRECTION: TRAINING-FREE SPATIAL CONSISTENCY INTERVENTION FOR DETR-BASED SMALL OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Small objects often occupy only a few visual tokens, making their localization unstable under changes in view or spatial context. We propose Frozen Spatial Consistency Intervention (FSCI), a training-free framework that turns this instability into a signal for correcting predictions from frozen detectors. Counterfactual equivariant explanation compares predictions under a horizontal flip to identify localization-sensitive instances and guide consistency-based correction. Nearest-center ownership slicing recovers local detail while assigning each candidate to a unique spatial source, reducing duplicates and candidate competition. Across five frozen DETR variants and three datasets,COCO, TinyPerson, and VisDrone2019—FSCI improves AP in all 15 detector–dataset settings without updating model weights. Diagnostic experiments further support the directional and spatial faithfulness of its explanations. These results show that spatial inconsistency can provide an actionable signal for inference-time detection.

1 INTRODUCTION

Small-object detection is fundamentally difficult because only a few visual tokens are available to support both recognition and precise localization. Consequently, minor changes in sampling, transformation, or spatial context can produce substantial changes in the predicted bounding box. This instability is particularly evident in Detection Transformer (DETR)-based detectors, where weak object evidence must also compete for a limited prediction budget. We therefore investigate a central question: can the spatial inconsistency exposed by a frozen detector be converted into supervision for correcting its own small-object predictions?

Most existing approaches treat small-object errors as a deficiency that must be compensated for externally. Training-time methods improve feature resolution, query formulation, or optimization objectives, but require additional training and annotations. Test-time methods recover local detail through multiview inference or image slicing, yet their gains often depend on black-box aggregation and may introduce duplicate candidates. Meanwhile, detection explanations usually visualize influential regions without providing an executable correction rule. Although these approaches differ in implementation, they share a common limitation: the frozen detector is treated as a passive prediction function rather than a system whose failure behavior can reveal how its outputs should be corrected.

Our key hypothesis is that the spatial knowledge required for correction is already encoded in the frozen detector but becomes observable only under controlled interventions. After inverse coordinate mapping, an image and its semantics-preserving horizontal flip should produce geometrically consistent detections. Similarly, when an object appears in multiple overlapping crops, the prediction obtained near the center of a crop should provide more reliable spatial evidence than predictions close to crop boundaries. Violations of these relations expose instance-level localization drift and crop-source conflict.

As illustrated in Fig. 1, we interpret these violations as spatial explanations generated by the detector itself. Unlike conventional test-time augmentation, the transformed views are used as diagnostic probes whose responses directly determine the subsequent intervention.

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.

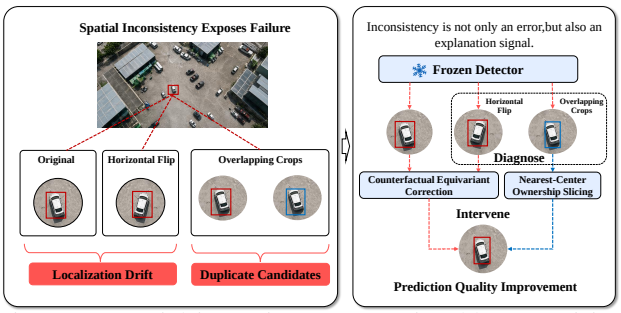

Figure 1: Spatial inconsistency as actionable supervision. Controlled horizontal flipping and overlapping crops expose localization drift and conflicting candidate sources in a frozen small-object detector. FSCI converts these violations into counterfactual equivariant correction and nearest-center ownership, forming an explain-and-correct loop without parameter updates.

Based on this principle, we propose Frozen Spatial Consistency Intervention (FSCI), a training-free framework that closes the loop from explanation to correction. FSCI first establishes same-class correspondence between the original and horizontally flipped views. The mapped cross-view differences reveal both whether a prediction is spatially unstable and in which direction it should be corrected. FSCI then performs counterfactual equivariant fusion for sufficiently supported instances. To recover fine-grained evidence, overlapping local crops are introduced. Their candidates are regulated by nearest-center ownership, which assigns each spatial location to a unique crop and suppresses redundant explanations before final prediction aggregation. The two components address complementary failures. Counterfactual intervention corrects localization drift, while ownership slicing reduces duplicate candidates and competition for the limited detection budget.

Extensive experiments on five frozen DETR variants, including DETR, Deformable DETR, Conditional DETR, DAB-DETR, and Anchor-DETR, evaluate FSCI on COCO, TinyPerson, and VisDrone2019. Under the same paired evaluation protocol, FSCI improves AP in all 15 detector–dataset settings without updating any model parameter. Mechanistic analysis on the COCO further shows that small objects exhibit substantially greater cross-view inconsistency and decoder-attention drift than large objects. Direction-reversal, random-intervention, and budget-matched ownership controls confirm that the gains depend on the correction direction and spatial ownership encoded by the explanations. Finally, an independently held-out translation intervention reduces spatial inconsistency on 66.36% of images, closing the loop from observed failure and internal explanation to correction and improved post-intervention behavior.

Our contributions are summarized as follows:

- We formulate frozen small-object detection improvement as explanation-guided spatial intervention, in which cross-view inconsistency serves simultaneously as a diagnostic signal and a source of correction.
- We introduce counterfactual equivariant explanation and nearest-center ownership slicing. Together, they correct localization-sensitive predictions, recover local details, and reduce duplicate candidate competition without training or fine-tuning.
- We validate FSCI on five frozen DETR variants across COCO, TinyPerson, and VisDrone2019, yielding 15 paired detector–dataset settings. FSCI improves AP in every setting. Ablation and interpretability experiments further verify the effectiveness and the directional and spatial faithfulness of the proposed interventions.

2 Related Work

2.1 Small-Object Detection

Small-object detection methods mainly improve training-time representations or recover image details at test time. Training-time approaches strengthen multiscale features, loss functions, query localization, and attention or anchor design (Lin et al., 2017a;b; Singh & Davis, 2018; Singh et al., 2018; Liu et al., 2018; Ghiasi et al., 2019; Pang et al., 2019; Tan et al., 2020; Yang et al., 2022; Zhu et al., 2021; Meng et al., 2021; Liu et al., 2022; Zhang et al., 2023). Although effective, they generally require retraining, annotations, or architecture-specific modifications. Test-time methods instead aggregate transformed predictions, fuse or suppress boxes, and perform sliced infer-

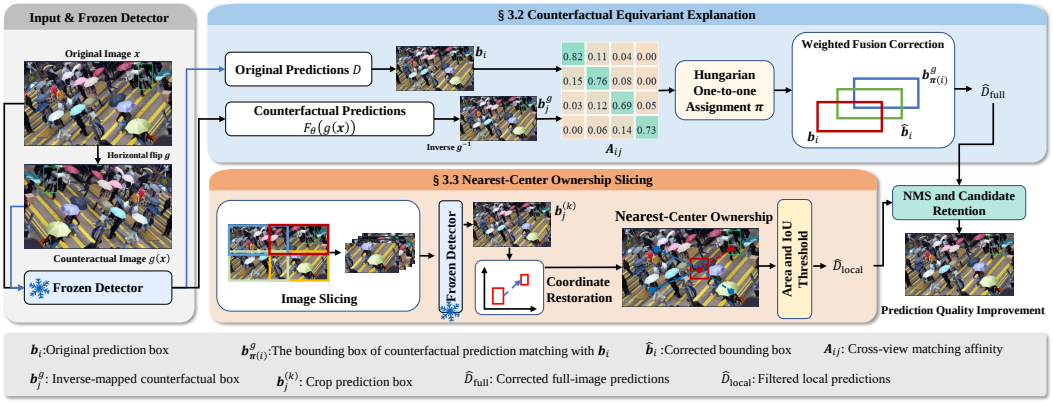

Figure 2: Overview of FSCI. A frozen detector processes the original and flipped images. After inverse mapping, class-constrained matching and confidence-weighted fusion produce corrected full-image predictions. Slice predictions are mapped back, filtered by ownership and geometry, and combined with full-image predictions via class-aware NMS and candidate retention. No model parameters are updated.

ence (Shanmugam et al., 2021; Solovyev et al., 2021; Bodla et al., 2017; Akyon et al., 2022). These methods preserve frozen models, but repeated views and overlapping crops generate duplicate predictions, increase dependence on non-maximum suppression, and intensify competition within a limited top-k budget. In contrast, our method interprets cross-view inconsistency and crop-source conflicts as spatial evidence, then performs traceable instance correction and unique ownership assignment without training.

2.2 Detection Explanation and Counterfactual Reasoning

Detection explanations can be broadly divided into attribution-based and counterfactual approaches. Attribution methods use gradients, path integration, local surrogates, game-theoretic attribution, or occlusion to identify influential image regions (Selvaraju et al., 2017; Chattopadhay et al., 2018; Sundararajan et al., 2017; Bach et al., 2015; Shrikumar et al., 2017; Smilkov et al., 2017; Ribeiro et al., 2016; Lundberg & Lee, 2017; Petsiuk et al., 2018; Fong & Vedaldi, 2017; Fong et al., 2019). Instance-aware methods and Transformer relevance propagation further adapt attribution to object detection (Petsiuk et al., 2021; Zhao & Chan, 2023; Chefer et al., 2021). However, most produce post-hoc saliency maps without an executable mapping to localization correction, while attention alone may not provide faithful explanations (Abnar & Zuidema, 2020; Jain & Wallace, 2019; Adebayo et al., 2018; Sixt et al., 2020). Counterfactual methods reveal decision evidence by deleting, replacing, or generating semantic content (Goyal et al., 2019; Wang & Vasconcelos, 2020; Zemni et al., 2023; Jeanneret et al., 2023; Dhurandhar et al., 2018), but often require generators or iterative optimization that entangle semantics, realism, and geometry. We instead employ an invertible and semantics-preserving horizontal flip. After inverse mapping and one-to-one instance matching, cross-view differences quantify localization risk and specify a correction direction. The explanation directly controls equivariant correction and nearest-center ownership, thereby closing the loop from diagnosis to intervention.

3 Method

3.1 Problem Formulation and Overall Architecture

Given an input image $\boldsymbol{x}$ and a parameter-frozen DETR-based detector F_θ, its prediction set is

$$\mathcal{D} = F_\theta(x) = \{(\boldsymbol{b}_i, c_i, s_i)\}_{i=1}^{N}, \tag{1}$$

where $\boldsymbol{b}_i$, c_i, and s_i denote the bounding box, class, and confidence score of the i-th prediction, respectively. Our objective is to use the model's own spatial inconsistency to generate an explanation $\mathcal{E}$ and obtain corrected detections without updating parameters, computing gradients, or using target-domain annotations: $\widehat{\mathcal{D}} = \psi(\boldsymbol{x}, F_\theta, \mathcal{E})$, where ψ denotes the inference-time intervention. As

shown in Fig. 2, FSCI consists of counterfactual equivariant explanation and nearest-center ownership slicing. First, a semantics-preserving counterfactual view exposes cross-view localization drift. Instance matching and confidence-weighted fusion then produce corrected full-image predictions. Next, overlapping slices recover local pixels, and nearest-center ownership assigns each local candidate to a unique spatial source before class-aware fusion with the full-image predictions. The procedure preserves model weights, the number of queries, and class semantics. It uses only spatial inconsistencies generated by the frozen model to perform deterministic inference-time adjustment.

3.2 Counterfactual Equivariant Explanation

A small object occupies only a few visual tokens, which makes its localization vulnerable to resampling and spatial transformations. Under a semantics-preserving horizontal flip, a detector should produce the same instances after inverse coordinate mapping. A substantial displacement of the same object across the two views therefore reveals localization instability in the frozen model. This module uses cross-view prediction differences to construct instance-level explanations and to provide an explicit direction for subsequent localization correction.

Specifically, given a horizontal-flip transformation g, we run inference on the original and flipped images and map the predictions from the flipped view back to the original coordinate system: $\mathcal{D}^g = g^{-1}(F_\theta(g(\boldsymbol{x})))$. Because query indices need not remain consistent across the transformation, we construct an affinity matrix using class consistency and bounding-box intersection over union (IoU):

$$\boldsymbol{A}_{ij} = \begin{cases} \mathrm{IoU}(\boldsymbol{b}_i, \boldsymbol{b}_j^g), & c_i = c_j^g, \\ -\infty, & \text{otherwise.} \end{cases} \tag{2}$$

We then use the Hungarian algorithm (Kuhn, 1955) to obtain a one-to-one assignment π. When $\boldsymbol{A}_{i,\pi(i)} \geq \tau$, prediction i is considered counterfactually supported. Here, τ is the counterfactual matching IoU threshold that determines whether predictions from the original and flipped views correspond to the same object. The index $\pi(i)$ denotes the prediction in the flipped view assigned to the i-th original prediction by Hungarian matching. The affinity $\boldsymbol{A}_{i,\pi(i)}$ measures the match between the two predictions. The counterfactual equivariant explanation for the i-th prediction is $e_i = (m_i, \rho_i, \Delta\boldsymbol{b}_i)$, where $m_i = \mathbb{1}[\boldsymbol{A}_{i,\pi(i)} \geq \tau]$ denotes the matching state, $\rho_i = \boldsymbol{A}_{i,\pi(i)}$ denotes cross-view spatial consistency, and $\Delta\boldsymbol{b}_i = \boldsymbol{b}^g_{\pi(i)} - \boldsymbol{b}_i$ denotes the localization displacement. A lower ρ_i or a larger $\|\Delta\boldsymbol{b}_i\|$ indicates greater sensitivity to spatial perturbations. A lower ρ_i or a larger $\|\Delta\boldsymbol{b}_i\|$ indicates greater sensitivity to spatial perturbations. Unlike methods that produce only saliency heatmaps, this explanation is defined directly in bounding-box coordinate space and therefore captures both localization risk and a potential correction direction.

To convert the explanation into an executable inference-time intervention, we apply equivariant fusion only to instances with valid counterfactual support. The corrected bounding box is defined as

$$\widetilde{\boldsymbol{b}}_i = \alpha_i \boldsymbol{b}_i + (1 - \alpha_i)\boldsymbol{b}^g_{\pi(i)}, \tag{3}$$

where

$$\alpha_i = \frac{s_i}{s_i + s^g_{\pi(i)} + \varepsilon}. \tag{4}$$

Here, s_i and $s^g_{\pi(i)}$ are the class confidences of the corresponding predictions in the original and counterfactual views, respectively. The constant ε is a small positive value that prevents division by zero. According to the counterfactual support state, we apply equivariant correction to matched instances and retain the original predictions without valid support. This yields the corrected full-image prediction set

$$\widehat{\mathcal{D}}_{\text{full}} = \left\{ (\widetilde{\boldsymbol{b}}_i, c_i, s_i) \right\}_{i=1}^{N}. \tag{5}$$

3.3 Nearest-Center Ownership Slicing

Full-image resizing substantially reduces the effective pixels of small objects, whereas local slicing restores spatial detail. Overlapping slices, however, cause the same object to appear repeatedly in multiple local views. These duplicate candidates increase false positives and compete with full-image predictions under non-maximum suppression (NMS) and the top-k output budget. Because

Table 1: Comparison on dataset COCO2017.

Method	AP	AP_{50}	AP_{75}	AP_s
DETR (Carion et al., 2020)	41.998	62.329	44.288	20.701
DETR + FSCI	42.729 (+0.731)	63.660 (+1.331)	44.780 (+0.492)	22.753 (+2.052)
Deformable DETR (Zhu et al., 2021)	43.721	62.784	47.563	24.363
Deformable DETR + FSCI	44.564 (+0.843)	64.053 (+1.269)	48.313 (+0.750)	26.098 (+1.735)
Conditional DETR (Meng et al., 2021)	40.874	61.859	43.262	20.542
Conditional DETR + FSCI	41.722 (+0.848)	63.354 (+1.495)	44.192 (+0.930)	22.330 (+1.788)
Anchor-DETR (Wang et al., 2022)	36.862	59.955	37.428	13.973
Anchor-DETR + FSCI	38.514 (+1.652)	62.052 (+2.098)	39.378 (+1.950)	18.251 (+4.277)
DAB-DETR (Liu et al., 2022)	40.399	61.463	42.299	18.505
DAB-DETR + FSCI	41.372 (+0.973)	63.207 (+1.744)	43.470 (+1.171)	21.507 (+3.002)

Table 2: Comparison on small target dataset TinyPerson.

Method	AP	AP_{50}	AP_{75}	AP_s
DETR (Carion et al., 2020)	3.480	10.619	1.556	1.248
DETR + FSCI	5.931 (+2.451)	18.262 (+7.643)	2.457 (+0.901)	4.050 (+2.802)
Deformable DETR (Zhu et al., 2021)	6.316	18.324	2.863	3.541
Deformable DETR + FSCI	10.110 (+3.794)	28.578 (+10.254)	4.515 (+1.652)	8.057 (+4.516)
Conditional DETR (Meng et al., 2021)	3.466	10.655	1.616	1.007
Conditional DETR + FSCI	5.884 (+2.418)	18.530 (+7.875)	2.301 (+0.685)	4.127 (+3.120)
Anchor-DETR (Wang et al., 2022)	1.404	4.366	0.581	0.688
Anchor-DETR + FSCI	2.358 (+0.953)	8.128 (+3.761)	0.909 (+0.328)	1.198 (+0.509)
DAB-DETR (Liu et al., 2022)	3.396	10.507	1.593	1.352
DAB-DETR + FSCI	5.732 (+2.337)	18.218 (+7.711)	2.256 (+0.663)	3.868 (+2.516)

an object generally retains more complete context near a crop center, this module assigns every local candidate to a unique and more reliable source crop.

Specifically, we partition the input image into a set of overlapping local crops $\mathcal{C} = \{C_k\}_{k=1}^{K}$, where k indexes a crop and K is the total number of crops. We use overlap ratio η in a $G \times G$ slicing grid, so $K = G \times G$. The detector processes each local view independently and maps every predicted box back to the original image coordinates: $\widetilde{\boldsymbol{b}}_j^{(k)} = T_k^{-1}\left(\boldsymbol{b}_j^{(k)}\right)$, where $\boldsymbol{b}_j^{(k)}$ denotes the j-th prediction from crop C_k, and T_k^{-1} denotes the inverse mapping from crop coordinates to the original image coordinates.

Let s denote the source-crop index of the current candidate. The point $q_j^{(s)}$ is the center of the globally mapped box $\widetilde{\boldsymbol{b}}_j^{(s)}$, and $\mathcal{K}(q_j^{(s)})$ is the set of crops covering that location. Nearest-center ownership is defined as

$$o\left(q_j^{(s)}\right) = \underset{k \in \mathcal{K}(q_j^{(s)})}{\arg\min} \left\| q_j^{(s)} - \boldsymbol{\mu}_k \right\|_2^2, \tag{6}$$

where $\boldsymbol{\mu}_k$ is the center of the k-th crop. We then determine whether source crop s owns the candidate location:

$$\mathrm{keep}\left(q_j^{(s)}, s\right) = \mathbb{1}\left[s = o\left(q_j^{(s)}\right)\right]. \tag{7}$$

Here, $\mathrm{keep}(\cdot)$ is the candidate-retention function, and $\mathbb{I}[\cdot]$ is the indicator function. This rule gives every location in an overlapping region a unique local explanation source, suppressing duplicate candidates introduced near crop boundaries. We retain only local predictions whose box area does not exceed $S_{\max}$. If a local candidate overlaps a same-class full-image prediction whose area exceeds $S_{\max}$ with an IoU of at least U, we remove the local candidate to prevent an unreliable small local box from suppressing a reliable large global box. The resulting local candidate set is $\widehat{\mathcal{D}}_{\mathrm{local}}$. We merge it with the full-image predictions using class-aware NMS:

$$\widehat{\mathcal{D}} = \mathrm{Top}_B\left(\mathrm{NMS_c}\left(\widehat{\mathcal{D}}_{\mathrm{full}} \cup \widehat{\mathcal{D}}_{\mathrm{local}}\right)\right), \tag{8}$$

where $\mathrm{NMS_c}$ denotes class-aware NMS and B is the final candidate budget. The full-image branch preserves global semantics and large-object detections, while the ownership-filtered local branch supplies complementary small-object evidence.

Table 3: Comparison on small target dataset VisDrone2019.

Method	AP	AP_{50}	AP_{75}	AP_s
DETR (Carion et al., 2020)	8.265	16.930	7.235	2.182
DETR + FSCI	9.455 (+1.190)	20.358 (+3.428)	7.546 (+0.311)	4.568 (+2.386)
Deformable DETR (Zhu et al., 2021)	11.116	20.752	10.307	3.894
Deformable DETR + FSCI	12.443 (+1.327)	24.347 (+3.595)	10.939 (+0.632)	6.795 (+2.901)
Conditional DETR (Meng et al., 2021)	8.764	17.360	7.812	1.946
Conditional DETR + FSCI	10.099 (+1.335)	21.532 (+4.172)	8.424 (+0.612)	4.772 (+2.826)
Anchor-DETR (Wang et al., 2022)	6.384	14.013	5.317	1.028
Anchor-DETR + FSCI	7.878 (+1.493)	18.340 (+4.327)	5.829 (+0.511)	3.239 (+2.211)
DAB-DETR (Liu et al., 2022)	8.190	16.384	7.203	1.688
DAB-DETR + FSCI	9.069 (+0.879)	19.400 (+3.016)	7.248 (+0.044)	3.710 (+2.022)

4 EXPERIMENTS

Datasets. We evaluate our method on COCO val2017 (Lin et al., 2014), TinyPerson test (Yu et al., 2020), and VisDrone2019-DET val (Zhu et al., 2022). COCO val2017 contains 5,000 images from 80 categories and supports a complete in-domain evaluation. TinyPerson test contains 786 high-resolution images. After merging `sea_person` and `earth_person` into `person` and excluding `uncertain`, `ignore`, and `logo` regions, we retain 13,687 person annotations. VisDrone2019-DET val contains 548 images. We map `pedestrian` and `people` to `person`, `van` to `car`, and `motor` to `motorcycle`. `tricycle` and `awning-tricycle` are excluded because they have no unambiguous COCO counterparts. For TinyPerson and VisDrone2019, we directly use frozen COCO-pretrained weights to assess the cross-domain generalization of the training-free intervention.

Evaluation metrics. On COCO, we use the official bounding-box metrics and report average precision (AP), AP at IoU thresholds of 0.50 and 0.75 (AP_{50} and AP_{75}), and AP for small objects (AP_s). After category mapping, TinyPerson and VisDrone2019 use the same COCO-style evaluator to ensure paired comparisons between each frozen baseline and FSCI. The TinyPerson results are not based on its official miss-rate metric, and the VisDrone2019 results do not follow the challenge AP protocol. In addition, $\mathrm{maxDets} = 100$ lowers absolute scores in dense scenes. We therefore focus the cross-domain evaluation on relative changes from each frozen baseline under the same protocol rather than on absolute comparisons with target-domain-specific training methods.

Implementation details. We evaluate FSCI on five frozen ResNet-50-based detectors—DETR, Deformable DETR, Conditional DETR, DAB-DETR, and Anchor-DETR—using their public COCO-pretrained weights. Only model-specific preprocessing and decoding adapters are used; all FSCI hyperparameters are shared across detectors and datasets. Following the original evaluation settings, the first three detectors resize the shorter side to 800 pixels with a maximum longer side of 1,333 pixels, while DAB-DETR and Anchor-DETR use 640 and 1,067 pixels, respectively. Each image is processed as the original view, its horizontal flip, and four overlapping crops ($G = 2$, 20% overlap). We set $S_{\max} = 1024\,\mathrm{px}^2$, accept same-class Hungarian matches with IoU ≥ 0.3, apply nearest-center ownership and class-wise NMS at 0.6, and retain at most 100 detections. Baselines use original-view predictions from the same cached runs. No target-domain annotations, gradients, training, score calibration, or additional output budget are used. Experiments run on one NVIDIA RTX 3060 GPU.

4.1 COMPARISON WITH FROZEN BASELINES

As shown in Tables 1–3, FSCI improves all five frozen detectors across COCO, TinyPerson, and VisDrone2019, covering all 15 detector–dataset combinations without target-domain annotations or parameter updates. On COCO, it yields an average gain of 1.009 AP and 2.571 AP_s, with Anchor-DETR achieving the largest AP improvement of 1.652 points. On TinyPerson, FSCI improves AP by 0.953–3.794 points (2.391 on average), with Deformable DETR obtaining the largest gains in both AP and AP_s. On VisDrone2019, it improves AP by 0.879–1.493 points (1.245 on average) and consistently raises AP_{50} and AP_s. These results demonstrate that FSCI generalizes across detector architectures and data distributions, particularly improving small-object recovery under domain shift.

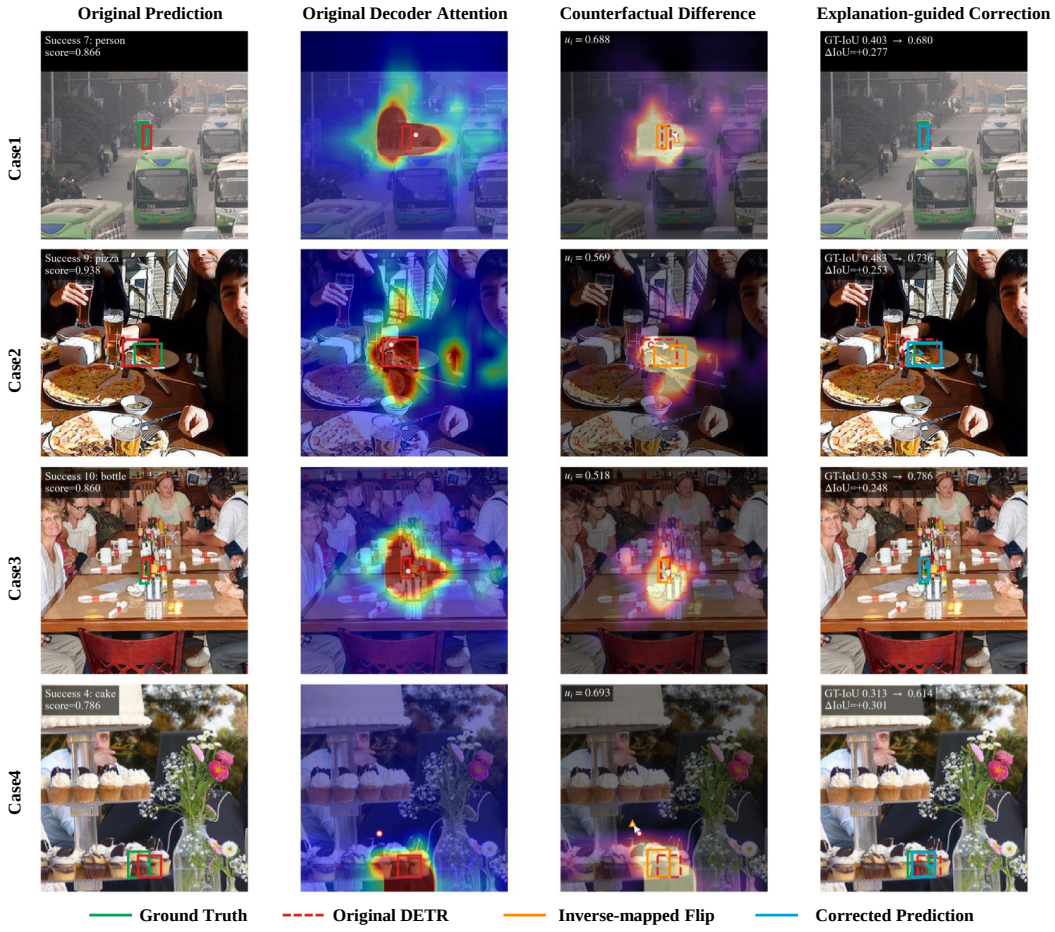

Figure 3: Counterfactual corrections on COCO. Rows show the original prediction, cross-attention, cross-view attention difference, and corrected prediction. Circle–triangle pairs indicate attention displacement. Green, dashed red, orange, and cyan boxes denote ground truth, original, flipped-view, and corrected predictions. Values report inconsistency and ΔGT-IoU; ground truth is visualization-only.

4.2 Ablation Study

Component ablation. Table 4 evaluates each component on COCO using identical cached DETR-R50 outputs. Counterfactual fusion and ownership slicing independently improve AP by 0.536 and 0.409 points, respectively, whereas combining fusion with naive slicing yields only a 0.009-point gain due to duplicate boundary predictions and candidate competition. With nearest-center ownership and large-box protection, the full method reaches 42.729 AP, improving AP and AP_s by 0.731 and 2.052 points. These results confirm that effective global–local integration requires spatial ownership and scale-aware protection.

Table 4: Component ablation of DETR-R50 on COCO val2017.

Variant	AP	ΔAP	ΔAP_s
Frozen baseline	41.998	+0.000	+0.000
Counterfactual fusion only	42.534	+0.536	+0.783
Ownership slicing only	42.407	+0.409	+1.718
Flip + naive slicing	42.007	+0.009	+1.052
+ nearest-center ownership	**42.729**	+0.731	+2.052

4.3 Instance-level visualization.

Figure 3 illustrates how FSCI converts counterfactual explanations into query-level localization corrections. Although the original DETR queries predict the correct categories, their boxes are spatially

misaligned. After horizontal flipping and inverse mapping, shifts in attention barycenters and predicted boxes reveal cross-view localization instability. FSCI uses the associated box displacement as the correction direction and performs confidence-weighted fusion. The corrected boxes achieve higher GT-IoU in all four examples, linking attention instability to actionable localization correction.

4.4 Mechanistic Interpretability and Analysis

We examine whether FSCI provides a mechanistically faithful explanation rather than merely a beneficial inference heuristic. The analysis follows a closed chain from observable failure and internal spatial evidence to directional intervention and held-out validation. All statistics are computed over the complete COCO val2017 set using frozen DETR-R50. Ground-truth annotations are accessed only after inference for diagnosis and evaluation, and the held-out transformation is never used to construct the correction.

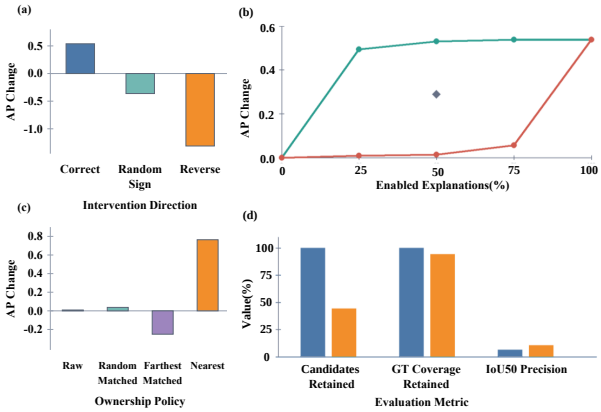

Figure 4: COCO explanations. (a) Directional faithfulness: correct, random-sign, and reversed displacement. (b) Selectivity: green activates high-to-low salience, red activates low-to-high salience, and the gray diamond is random 50%. (c) Budget-matched ownership: raw, random, farthest-center, and nearest-center policies. (d) Efficiency: raw slicing (blue) versus nearest ownership (orange) for candidate retention, retained GT coverage, and IoU50 precision. GT is diagnostic-only.

Directional faithfulness and dose response. As shown in Fig. 4(a), applying the confidence-weighted counterfactual displacement in the direction indicated by the explanation improves DETR by 0.538 AP. Reversing the same displacement decreases AP by 1.311 points, while five random-sign trials produce an average decrease of 0.362 ± 0.043 points. All controls preserve displacement magnitude, matched instances, classes, scores, and candidate sets. Figure 6(d) further varies the correction strength λ. Performance peaks at $\lambda = 1.0$ and decreases at $\lambda = 1.5$, revealing an over-correction regime. The explanation therefore specifies an actionable direction and magnitude rather than inducing generic box smoothing.

Diagnostic selectivity. Fig. 4(b) compares activation strategies after ranking explanations by salience. Activating only the top 25% most salient explanations yields a 0.494-point AP gain, accounting for 91.7% of the full correction gain, whereas the bottom 25% yields only 0.008 points. Randomly activating 50% gives 0.288 ± 0.022 points, substantially below the 0.530-point gain obtained from the top 50%. Across 70,221 diagnostic instances, cross-view inconsistency correlates with localization error at $\rho = 0.550$. The corresponding image-level correlation is $\rho = 0.806$ (95% confidence interval (CI) [0.793, 0.818]), and Fig. 6(c) shows a nearly monotonic increase in localization error across inconsistency deciles. In addition, the highest-salience quartile achieves a mean GT-IoU



Figure 5: Scale-dependent instability of DETR-R50 on COCO val2017. Bars show cross-view inconsistency, localization error, and normalized attention drift across object scales. Instance counts are shown below each category. Ground truth is diagnostic-only.

correction 15.8 times that of the lowest quartile. These results establish that the explanation ranks localization risk rather than simply selecting high-confidence predictions.

Spatial faithfulness and candidate efficiency. As shown in Fig. 4(c), nearest-center ownership improves AP by 0.764 points. Five random-ownership trials with the same per-image candidate budget yield only 0.037 ± 0.051 points, while equal-budget farthest-center ownership decreases AP by 0.251 points. Fig. 6(e) provides the underlying spatial evidence: crop predictions are more reliable in crop interiors and deteriorate near crop boundaries. Accordingly, retained candidates have a smaller normalized source-center distance than rejected candidates (0.267 versus 0.381). At score ≥ 0.05, nearest-center ownership retains 44.2% of the original crop candidates while preserving 94.33% of the ground-truth objects covered by the original slices. As shown in Fig. 4(d), candidate precision at IoU 0.50 increases from 6.11% to 10.32%, and valid candidates per covered ground-truth object decrease from 1.741 to 1.378. The gain therefore follows the measured spatial reliability distribution rather than candidate pruning alone.

Scale-dependent internal mechanism. As shown in Fig. 6(a) and Fig. 5, small objects exhibit substantially greater spatial instability than medium and large objects. The distribution in Fig. 6(a) shifts progressively toward lower cross-view inconsistency as object scale increases. Consistently, the mean inconsistency decreases from 0.3055 for small objects to 0.1176 for large objects. Fig. 5 further shows that localization error decreases from 0.5090 to 0.3220, while final-layer attention drift drops more sharply from 0.3045 to 0.0534. These consistent output- and representation-level trends connect small-object localization failures to unstable internal spatial evidence.

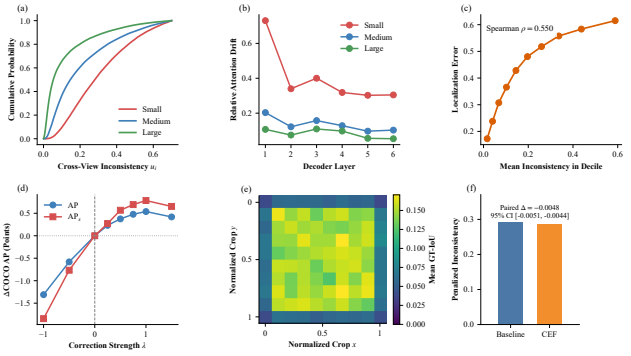

Figure 6: Mechanistic analysis of frozen DETR-R50 on COCO val2017. (a) Cross-view inconsistency across object scales. (b) Attention drift across decoder layers. (c) Localization error by inconsistency decile. (d) Correction-strength response. (e) Spatial reliability of crop predictions. (f) Inconsistency under a held-out translation. Ground truth is diagnostic-only.

Held-out interventional closure. We evaluate whether corrections derived from horizontal reflection transfer to an unseen four-pixel translation. After inverse mapping, same-class predictions are matched by Hungarian assignment, and inconsistency is measured as mean $1 - \mathrm{IoU}$ with unit cost for unmatched predictions. As shown in Fig. 6(f), FSCI reduces inconsistency from 0.291811 to 0.287059 ($\Delta = -0.004752$, 95% CI $[-0.005116, -0.004391]$), improving 66.36% of images. AP and $\mathrm{AP_s}$ also increase by 0.464 and 0.644 points. Since translation is not used to construct the correction, these results provide non-circular interventional validation.

Boundary of the explanation. Cross-view inconsistency is a reliable indicator of localization risk, but it is not by itself a monotonic predictor of correction gain. Its instance-level correlation with the IoU gain after correction is -0.004, and the most unstable decile exhibits mild over-correction. FSCI therefore combines correspondence validity, confidence, and relative displacement rather than applying an unconditional threshold on inconsistency. This boundary result narrows the causal claim and distinguishes diagnosis from intervention eligibility.

5 Conclusion

We presented FSCI, a training-free framework that converts spatial inconsistency into executable explanations. Counterfactual equivariant correction and nearest-center ownership slicing mitigate localization drift and candidate competition. Experiments across multiple datasets demonstrate consistent improvements in small-object detection and establish a transparent paradigm for correcting predictions through model responses. Future work will investigate adaptive counterfactual transformations and interventions on internal representations to improve generalization across architectures.

AI USE STATEMENT

Large language models (LLMs) were used solely to improve the writing of this paper, including grammar, clarity, and readability. They were not used to generate research ideas, design the methodology or experiments, conduct analyses, interpret results, or produce scientific claims. All AI-assisted revisions were reviewed by the authors, who take full responsibility for the final content of this work.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.385. URL https://aclanthology.org/2020.acl-main.385/.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Fatih Cagatay Akyon, Sinan Onur Altinuc, and Alptekin Temizel. Slicing aided hyper inference and fine-tuning for small object detection. In *IEEE International Conference on Image Processing*, 2022.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):e0130140, 2015.

Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-NMS: Improving object detection with one line of code. In *IEEE International Conference on Computer Vision*, 2017.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 2020.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 839–847, 2018.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 782–791, 2021.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2950–2958, 2019.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3429–3437, 2017.

Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V. Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7036–7045, 2019.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, volume 97, pp. 2376–2384. PMLR, 2019. URL https://proceedings.mlr.press/v97/goyal19a.html.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Adversarial counterfactual visual explanations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, 2017a.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, 2017b.

Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *International Conference on Learning Representations*, 2022.

Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8759–8768, 2018.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 2017.

Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional DETR for fast training convergence. In *IEEE/CVF International Conference on Computer Vision*, 2021.

Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 821–830, 2019.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference*, 2018.

Vitali Petsiuk, Rajiv Jain, Varun Manjunatha, Vlad I. Morariu, Ashutosh Mehra, Vicente Ordonez, and Kate Saenko. Black-box explanation of object detectors via saliency maps. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you? explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, 2017.

Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. Better aggregation in test-time augmentation. In *IEEE/CVF International Conference on Computer Vision*, 2021.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3145–3153, 2017.

Bharat Singh and Larry S. Davis. An analysis of scale invariance in object detection – snip. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3578–3587, 2018.

Bharat Singh, Mahyar Najibi, and Larry S. Davis. Sniper: Efficient multi-scale training. In *Advances in Neural Information Processing Systems*, volume 31, pp. 9310–9320, 2018.

Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified bp attributions fail. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9046–9057, 2020.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: Removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, 2021.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 2017.

Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10781–10790, 2020.

Pei Wang and Nuno Vasconcelos. SCOUT: Self-aware discriminant counterfactual explanations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor DETR: Query design for transformer-based detector. In *AAAI Conference on Artificial Intelligence*, 2022.

Chenhongyi Yang, Zehao Huang, and Naiyan Wang. Querydet: Cascaded sparse query for accelerating high-resolution small object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13668–13677, 2022.

Xuehui Yu, Yuqi Gong, Nan Jiang, Qixiang Ye, and Zhenjun Han. Scale match for tiny person detection. In *IEEE Winter Conference on Applications of Computer Vision*, 2020.

Mehdi Zemni, Mickaël Chen, Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez, and Matthieu Cord. OCTET: Object-aware counterfactual explanations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations*, 2023.

Chenyang Zhao and Antoni B. Chan. ODAM: Gradient-based instance-specific visual explanations for object detection. In *International Conference on Learning Representations*, 2023.

Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. volume 44, pp. 7380–7399, 2022. doi: 10.1109/TPAMI.2021.3119563.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

A Appendix