READING DECODER TRAJECTORIES: TRAINING-FREE COUNTERFACTUAL QUERY-TRAJECTORY RELIABILITY FOR SMALL-OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Small-object detection remains challenging because limited pixels cause information loss and suppress the scale knowledge encoded in pretrained detectors. Existing approaches mainly improve representations through multiscale training, architecture redesign, or parameter adaptation, implicitly assuming that frozen models lack the required capability. We challenge this assumption and hypothesize that small-object knowledge already exists in frozen detectors but remains underactivated and unstable during query evolution. To test this hypothesis, we propose Counterfactual Query-Trajectory Reliability (CQTR), a training-free framework that elicits latent responses through counterfactual scale interventions and interprets candidate reliability from decoder-internal spatial convergence, semantic persistence, and cross-scale conflicts. A small unlabeled training subset selects the appropriate correction mechanism for each model–data stream, without parameter updates or target-domain annotations. Across 27 combinations of nine frozen detectors and three datasets, CQTR consistently improves average precision (AP) and average precision for small objects (AP_s). Closed-loop analyses further show that scale intervention activates latent responses, trajectory evidence predicts ground-truth support, and unlabeled routing selects the more effective branch. CQTR therefore reframes small-object detection from external scale augmentation to the activation and reliability assessment of latent scale knowledge.

1 INTRODUCTION

Small-object detection is fundamental to visual perception systems such as autonomous driving, unmanned aerial vehicle perception, and remote-sensing analysis. Because a small object occupies only a few pixels and feature cells, its visual evidence is easily weakened by spatial downsampling and overwhelmed by background responses. Detecting such objects therefore requires not only sufficient feature resolution, but also the reliable formation of object hypotheses from limited internal evidence. This raises a fundamental question: must small-object capability be acquired through additional training, or is the required scale knowledge already encoded in a pretrained detector but insufficiently activated during standard inference?

Set-prediction detectors represented by the Detection Transformer (DETR) formulate object detection through learnable queries and global matching (Carion et al., 2020). Subsequent studies improve their query mechanisms through deformable sampling, dynamic anchors, query selection, and efficient encoding (Zhu et al., 2021; Liu et al., 2022; Zhang et al., 2023; Huang et al., 2025). Nevertheless, their performance on small objects remains substantially lower than that on larger objects. Existing approaches commonly address this gap through multiscale training, high-resolution features, image slicing, or parameter adaptation. These methods improve external representations but do not determine whether the resulting gains come from newly learned knowledge or the reactivation of capabilities already present in a frozen model. Meanwhile, existing interpretability methods mainly attribute final predictions to input regions or attention distributions. They provide limited insight into how queries progressively establish spatial and semantic decisions during decoding.

We argue that small-object knowledge does not completely disappear from frozen detectors. Instead, spatial compression may leave this knowledge latent or unstable during query evolution. Increas-

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.

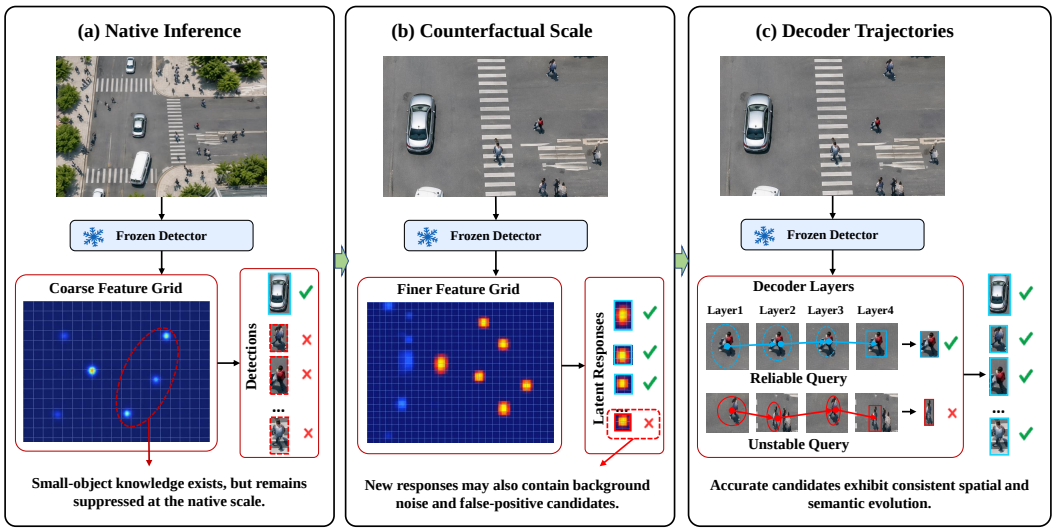

Figure 1: Motivation for CQTR. Counterfactual scaling activates latent small-object evidence in frozen detectors, while spatial-semantic decoder trajectories distinguish reliable queries from amplified noise.

ing the inference scale provides a counterfactual observation in which compressed object evidence can be reintroduced into the decoding process. However, the same intervention may also amplify background textures and spurious candidates. The central problem is therefore not simply to combine predictions from multiple scales, but to identify which scale-activated responses reflect reliable object knowledge by examining their internal query dynamics, as illustrated in Fig. 1.

To address this problem, we propose Counterfactual Query-Trajectory Reliability (CQTR), a training-free framework for activating and assessing latent scale knowledge in frozen detectors. CQTR constructs a counterfactual scale branch and interprets intermediate decoder states through spatial convergence and semantic persistence. It further detects unreliable candidates from semantic conflicts between native-scale and counterfactual predictions. Because query trajectories and cross-scale conflicts characterize different failure modes, CQTR uses a small unlabeled subset of the corresponding training split to select the appropriate correction mechanism for each model-data stream. The entire framework keeps the detector architecture and parameters unchanged, requires no backpropagation, and accesses no target-domain training annotations.

We evaluate CQTR across 27 combinations of nine frozen detectors and three datasets, including COCO2017, TinyPerson, and VisDrone2019. CQTR consistently improves AP, while AP_s increases in every setting. Closed-loop analyses show that counterfactual scaling activates latent small-object responses, query-trajectory evidence predicts candidate reliability, and unlabeled routing selects the more effective correction mechanism. Internal heatmaps and detection visualizations further connect changes in decoder states to improved small-object recall. These results support the view that small-object detection can be approached as the activation and reliability assessment of latent scale knowledge rather than solely as external scale augmentation. The main contributions are as follows:

- We reformulate small-object failure in frozen detectors as a problem of latent scale-knowledge activation and reliability assessment, distinguishing underactivated knowledge from genuinely missing capability.
- We introduce CQTR, a training-free framework that interprets and corrects scale-activated candidates through counterfactual intervention, spatial-semantic query trajectories, and cross-scale semantic conflicts.
- We conduct extensive cross-model and cross-domain experiments. Consistent improvements, closed-loop analyses, and internal visualizations establish an evidence chain from latent knowledge activation to query assessment and improved small-object detection.

2 Related Work

2.1 Small-Object Detection

Small-object detectors mainly enhance representations, decompose image space, or optimize object queries. Representation methods recover fine details through feature pyramids, high-resolution fusion, or super-resolution (Lin et al., 2017; Li et al., 2017; Bai et al., 2018; Ghiasi et al., 2019; Tan et al., 2020; Yang et al., 2022). Spatial methods enlarge relative object scale through scale normalization, slicing, or density-guided cropping (Singh & Davis, 2018; Singh et al., 2018; Najibi et al., 2019; Yang et al., 2019; Kisantal et al., 2019; Akyon et al., 2022). Query methods refine attention, reference points, initialization, and matching in set-prediction detectors (Carion et al., 2020; Zhu et al., 2021; Meng et al., 2021; Dai et al., 2021; Liu et al., 2022; Li et al., 2022; Zhang et al., 2023; Hou et al., 2024; Huang et al., 2025). These approaches generally address insufficient visible evidence by learning new scale representations or query priors, without examining whether frozen detectors already encode usable small-object knowledge. CQTR instead exposes latent responses through counterfactual scaling and evaluates their reliability during decoding.

2.2 Internal Model Interpretation

Model interpretation broadly includes post hoc attribution and internal-mechanism analysis. Attribution methods trace predictions to image regions using gradients, saliency, concepts, or attention propagation (Selvaraju et al., 2017; Sundararajan et al., 2017; Kim et al., 2018; Chefer et al., 2021b;a). They identify evidence supporting an output but reveal less about how it forms. Mechanistic methods instead examine hidden units, intermediate representations, and computational pathways (Bau et al., 2017; Khakzar et al., 2021; Dosovitskiy et al., 2021; Raghu et al., 2021). Detection studies further visualize attention or analyze query behavior (Carion et al., 2020; Zhu et al., 2021; Ye et al., 2023; Senthivel & Vu, 2024), yet rarely explain how scale changes affect query localization and semantics. CQTR reads decoder trajectories under counterfactual scaling and interprets reliability through spatial convergence, semantic persistence, and cross-scale conflict, shifting explanation from output attribution toward query-formation mechanisms.

3 Method

3.1 Problem Formulation and Overall Architecture

Given an input image $\boldsymbol{x}$ and a detector F_θ with frozen parameters, let S_0 denote the native preprocessing scale. The native branch produces D_0, where each candidate contains a bounding box, class label, and confidence score in the original-image coordinate system: $D_0 = F_\theta(T_{S_0}(\boldsymbol{x}))$. Here, T_{S_0} denotes the preprocessing transformation applied to the input image at native scale S_0. CQTR changes only the input scale to construct a counterfactual branch and retains box and class responses from the last several decoder layers. Trajectory interpretation and cross-scale conflict interpretation then correct the counterfactual candidates D_c in parallel, after which unlabeled routing selects the more reliable correction stream. We summarize the overall process as $\widehat{D} = \mathcal{F}(D_0, D_c)$. Here, $\mathcal{F}(\cdot)$ denotes the training-free correction process composed of query-trajectory interpretation, cross-scale semantic-conflict interpretation, unlabeled stream routing, and candidate fusion, as shown in Fig. 2.

3.2 Counterfactual Scale Branch

Native-scale resizing may compress small objects into only a few feature cells, while uniformly increasing the scale can amplify interpolation redundancy and background noise. CQTR therefore selects the counterfactual scale adaptively according to the input's spatial compression relative to the native scale.

Given the height H_x and width W_x of the original image, we define the spatial compression ratio as $\rho(\boldsymbol{x}) = \frac{\sqrt{H_x W_x}}{S_0}$. A larger ratio indicates stronger information compression under native resizing. The counterfactual scale is selected as

$$S(\boldsymbol{x}) = \begin{cases} S_1, & \rho(\boldsymbol{x}) \le \tau_p, \\ S_2, & \rho(\boldsymbol{x}) > \tau_p, \end{cases} \tag{1}$$

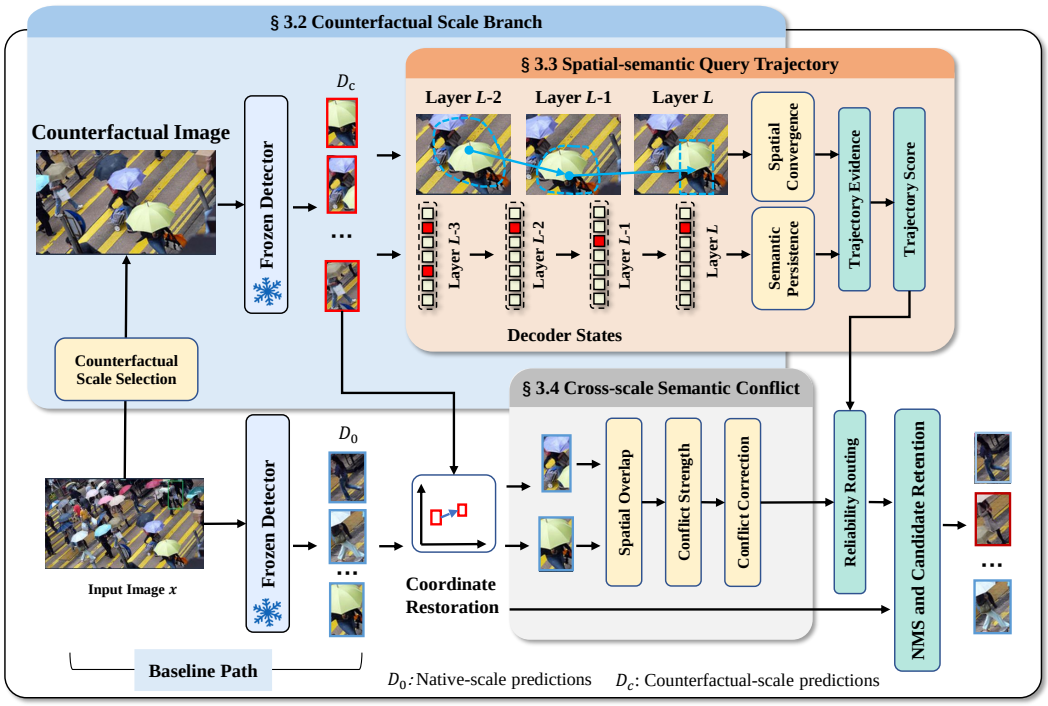

Figure 2: Overview of CQTR. CQTR interprets native and counterfactual predictions through decoder trajectories and cross-scale semantic conflicts, then uses unlabeled reliability routing to select the correction stream for final fusion. Snowflakes denote frozen parameters.

where S_1 and S_2 are the lower and higher counterfactual scales, respectively, and τ_p is the spatial-compression threshold. With detector parameters frozen, the counterfactual forward pass is $D_c = F_\theta\big(T_{S(\boldsymbol{x})}(\boldsymbol{x})\big)$. For a detector that dynamically derives spatial relations from the current feature shapes, we use its native forward pass directly. Otherwise, we save and temporarily clear the preset encoder and decoder inference sizes before this forward pass. This prompts the model to reconstruct positional embeddings, decoder anchors, and validity masks from the actual multiscale feature shapes. The original sizes are restored immediately afterward. Following spatial-semantic query-trajectory interpretation, counterfactual predictions are mapped back to the original-image coordinate system while the intermediate responses from the final decoder layers are retained.

3.3 Three-Step Spatial-Semantic Query-Trajectory Interpretation

Final confidence ignores how a prediction evolves during decoding. CQTR therefore estimates reliability from spatial convergence over the last three box transitions and semantic persistence over the final four decoder layers, then continuously recalibrates counterfactual scores.

Assume that the counterfactual branch contains L decoder layers. For final candidate i, let $b_i^{(l)}$ denote its normalized box at layer l, and let $p_{i,y}^{(l)}$ denote its class response. Its final predicted class is $y_i = \arg\max_y p_{i,y}^{(L)}$. First, we measure spatial convergence by the geometric mean of intersection over union (IoU) over the last three box transitions:

$$a_i = \exp\left[\frac{1}{3}\sum_{l=L-2}^{L}\log\Big(\max\Big(\mathrm{IoU}(\boldsymbol{b}_i^{(l-1)}, \boldsymbol{b}_i^{(l)}), \epsilon\Big)\Big)\right]. \tag{2}$$

Here, ϵ is a small constant that prevents numerical instability. A larger a_i indicates that the query box stabilizes toward the end of decoding, while a smaller a_i indicates persistent positional jumps.

Spatial convergence alone does not guarantee class reliability. To measure persistent support for the final class throughout decoding, we compare the final-class response with the strongest class

response in each of the last four layers:

$$c_i = \exp\left[\frac{1}{4}\sum_{l=L-3}^{L}\log\left(\max\left(\frac{p_{i,y_i}^{(l)}}{\max_y p_{i,y}^{(l)}}, \epsilon\right)\right)\right]. \tag{3}$$

When the final class remains close to the strongest class in consecutive decoder layers, c_i approaches 1. Frequent switching between classes reduces c_i. We then define query-trajectory evidence as the geometric mean of spatial convergence and semantic persistence: $e_i = \sqrt{a_i c_i}$. To avoid removing small objects with a fixed threshold, we recalibrate counterfactual confidence only relative to the median evidence of the current image:

$$s_i^{\text{traj}} = \text{clip}_{[0,1]}\left(s_i w_i^{\text{traj}}\right), \tag{4}$$

where

$$w_i^{\text{traj}} = \text{clip}_{[w_{\min}, w_{\max}]}(\exp[\beta\,(e_i - \text{med}(\boldsymbol{e}))]). \tag{5}$$

Here, s_i is the original confidence of counterfactual candidate i, $\boldsymbol{e}$ is the set of trajectory-evidence values for all counterfactual candidates in the current image, and β controls the correction strength. The operator $\text{clip}_{(a,b)}(\cdot)$ clips its input to the interval (a, b).

3.4 Cross-Scale Semantic-Conflict Interpretation

Stable trajectories may still produce inconsistent semantic explanations across scales because of context drift or class competition. To preserve newly exposed objects, CQTR soft-suppresses only counterfactual candidates that strongly overlap high-confidence native predictions but predict different classes.

For counterfactual candidate i, let $\boldsymbol{b}_i^c$, y_i^c, and s_i^c denote its box, class, and original confidence. The corresponding quantities for native candidate k are $\boldsymbol{b}_k^0$, y_k^0, and s_k^0. After both prediction streams are restored to the original-image coordinate system, we define confidence-weighted cross-scale spatial overlap as $w_{ik} = \text{IoU}(\boldsymbol{b}_i^c, \boldsymbol{b}_k^0)\sqrt{s_k^0}$. The maximum overlap with a native candidate of a different class defines the conflict strength: $r_i = \max_{k:\, y_k^0 \neq y_i^c} w_{ik}$. A larger conflict strength indicates a stronger contradiction between the two scale-specific interpretations of the same region. The conflict-corrected score is

$$s_i^{\text{conf}} = s_i^c \max(w_{\min}, 1 - \gamma r_i), \tag{6}$$

where γ controls the suppression strength and $w_{\min}$ is the minimum retention factor.

Trajectory recalibration and conflict suppression form parallel candidate streams and are not applied sequentially. The unlabeled router in the next subsection determines which stream is used.

3.5 Unlabeled Stream-Level Reliability Routing and Unified Output

Query trajectories and cross-scale semantic conflicts capture complementary forms of internal instability, and applying both corrections indiscriminately may suppress useful signals. CQTR therefore estimates reliability from a small unlabeled sample and selects trajectory recalibration or conflict suppression for the entire model-data stream.

For each model-data stream, we sample M unlabeled images from the training split using a fixed random seed. We first define stream-level trajectory dispersion as $\bar{d} = \frac{1}{M}\sum_{m=1}^{M}\text{Std}_i\left(e_i^{(m)}\right)$. Higher trajectory dispersion means that trajectory evidence more clearly separates candidate reliability. We define the conflict statistic as $\bar{r} = \frac{1}{M}\sum_{m=1}^{M}\max_i r_i^{(m)}$. A larger conflict statistic indicates stronger class competition caused by scale changes. Based on these statistics, we define the stream-level routing variable as $z = \mathbb{I}\left[\bar{d} \geq \tau_d \;\wedge\; \bar{r} \geq \tau_c\right]$, where τ_d and τ_c are the thresholds for trajectory dispersion and conflict strength. The operator $\mathbb{1}[\cdot]$ is the indicator function. We apply query-trajectory recalibration when $z = 1$. Otherwise, we apply cross-scale semantic-conflict suppression. After unlabeled calibration, the routing variable remains fixed for the entire model-data stream. The corrected counterfactual predictions are

$$\widetilde{D}_c = \begin{cases} \{(\boldsymbol{b}_i^c, y_i^c, s_i^{\text{traj}})\}_{i=1}^{Q}, & z = 1, \\ \{(\boldsymbol{b}_i^c, y_i^c, s_i^{\text{conf}})\}_{i=1}^{Q}, & z = 0. \end{cases} \tag{7}$$

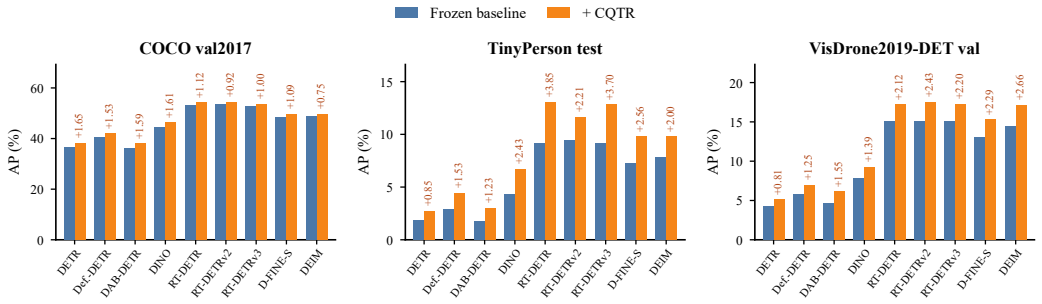

Figure 3: Overall AP comparison on COCO, TinyPerson, and VisDrone2019. Numbers above the orange bars denote absolute AP gains.

Finally, the spatial compression ratio $\rho(x)$ defined in Section 3.2 determines whether the native branch is retained:

$$\widehat{D} = \begin{cases} \mathrm{TopQ}\Big(\mathrm{ClassNMS}(D_0 \cup \widetilde{D}_c, \eta)\Big), & \rho(\boldsymbol{x}) \le \tau_r, \\ \widetilde{D}_c, & \rho(\boldsymbol{x}) > \tau_r, \end{cases} \tag{8}$$

where Q is the detector's candidate budget and η is the threshold for class-aware non-maximum suppression (NMS).

4 Experiments

Datasets. We evaluate on COCO val2017 (Lin et al., 2014), TinyPerson test (Yu et al., 2020), and VisDrone2019-DET val (Du et al., 2019). COCO contains 5,000 validation images. TinyPerson and VisDrone contain 786 and 548 evaluation images, respectively. For the latter two datasets, we directly apply COCO-pretrained weights without accessing target-domain training annotations or performing fine-tuning, thereby testing generalization to cross-domain small-object scenes.

Evaluation metrics. All three datasets use the same COCO-style bounding-box evaluator. We report AP averaged over IoU thresholds, AP at IoU thresholds of 0.50 and 0.75 (AP_{50} and AP_{75}), and AP_{s}. TinyPerson results do not follow its official miss-rate protocol. VisDrone results are also subject to the COCO-style limit of 100 maximum detections per image (maxDets=100). We therefore compare CQTR only with its paired frozen baseline under the same protocol and do not directly compare absolute values with published results obtained under different protocols.

Implementation details. We evaluate DETR-R50 (Carion et al., 2020), Deformable DETR-R50 (Zhu et al., 2021), DAB-DETR-R50 (Liu et al., 2022), DINO-R50 (Zhang et al., 2023), RT-DETR-R50 (Zhao et al., 2024), RT-DETRv2-R50 (Lv et al., 2024), RT-DETRv3-R50 (Wang et al., 2025), D-FINE-S (Peng et al., 2025), and DEIM-D-FINE-S (Huang et al., 2025). These models cover fixed queries, dynamic anchors, deformable sampling, query selection, and real-time hybrid encoding. Every model uses COCO-pretrained weights and runs with frozen parameters under inference mode without gradients. TinyPerson and VisDrone experiments use no target-domain training annotations and perform no fine-tuning. We use S0/S1/S2=640/800/960, $\tau_p = 2.0$, $\tau_r = 1.0$, an NMS threshold of 0.8, and at most 300 internal candidates. The trajectory interpretation uses the last three box transitions and the last four class responses with $\beta = 0.05$. The semantic-conflict coefficient is $\gamma = 0.10$. For each model-dataset combination, we use random seed 31415 to sample 100 unlabeled images from the corresponding training split for one-time stream calibration, with $\tau_d = 0.22$ and $\tau_c = 0.85$. On COCO, calibration routes RT-DETR-R50, RT-DETRv3-R50, and DEIM-D-FINE-S to query-trajectory recalibration and the remaining models to cross-scale semantic-conflict suppression. Every model on TinyPerson and VisDrone is routed to the conflict branch. The selected route remains fixed throughout full evaluation. We use the same configuration for all nine detectors and three datasets. Full evaluation covers 5,000, 786, and 548 images, respectively, for a total of 27 model-dataset combinations. All experiments run on one NVIDIA RTX 3060 GPU.

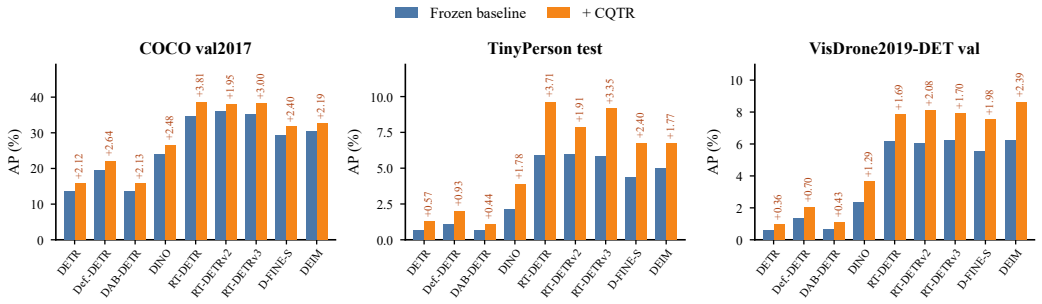

Figure 4: Small-object AP comparison on COCO, TinyPerson, and VisDrone2019. Numbers above the orange bars denote absolute improvements over the corresponding frozen baselines.

Table 1: Component ablation of RT-DETR-R50 across the three datasets.

Configuration	COCO AP	COCO AP_s	Tiny AP	Tiny AP_s	Vis AP	Vis AP_s
Frozen baseline	53.09	34.74	9.19	5.89	15.11	6.16
Counterfactual scale only	54.18	38.51	12.99	9.41	17.13	7.88
+ Trajectory recalibration	54.21	38.55	12.96	9.39	17.12	7.86
+ Conflict suppression	54.16	38.44	13.04	9.60	17.23	7.85
CQTR stream routing	54.21	38.55	13.04	9.60	17.23	7.85

4.1 Results Across Detectors and Datasets

Fig. 3 and 4 compare CQTR with nine frozen detectors on COCO, TinyPerson, and VisDrone2019. The figures visualize overall AP and AP_s, respectively. Complete numerical results for AP, AP_{50}, AP_{75}, and AP_s are reported in Tables 2–4 of Appendix A.1.

As shown in Figs. 3 and 4 and detailed in Tables 2–4, CQTR improves both AP and AP_s across all 27 detector-dataset combinations. On COCO, the gains reach 0.75–1.65 AP and 1.95–3.81 AP_s, with RT-DETR-R50 improving from 53.09 to 54.21 AP and from 34.74 to 38.55 AP_s. On TinyPerson, CQTR yields gains of 0.85–3.85 AP and 0.44–3.71 AP_s, where RT-DETR-R50 increases from 9.19 to 13.04 AP. On VisDrone2019, the improvements range from 0.81 to 2.66 AP and from 0.36 to 2.39 AP_s. DEIM-D-FINE-S achieves the largest AP gain, rising from 14.50 to 17.16, while RT-DETRv2-R50 obtains the highest absolute AP of 17.49. These consistent improvements across diverse detector families and target domains demonstrate that CQTR rec is neither architecture architecture-specific nor limited to the COCO scale distribution.

4.2 Ablation Study

Table 1 decomposes CQTR on RT-DETR-R50 into counterfactual scaling, trajectory recalibration, conflict suppression, and stream routing. Counterfactual scaling alone contributes most of the gain, reaching 54.18, 12.99, and 17.13 AP on COCO, TinyPerson, and VisDrone, respectively. Under overall AP, the trajectory rule performs best on COCO but is slightly below scale-only inference on TinyPerson and VisDrone. Conflict suppression performs better on the two cross-domain datasets but slightly decreases COCO AP. Using unlabeled statistics to select the corresponding branch, stream routing improves scale-only inference by 0.03, 0.05, and 0.10 AP, respectively. These results explain why the later components should not be applied unconditionally. They are weak ranking corrections for different internal failure modes, rather than additional scale enhancers.

4.3 Qualitative Results

As shown in Fig. 5, CQTR recovers small objects missed by the frozen baseline on all three datasets under a shared confidence threshold. The numbers of true positives in the COCO, TinyPerson, and VisDrone examples increase from 18, 9, and 43 to 23, 28, and 61, while false negatives decrease from 18, 68, and 131 to 13, 49, and 113. False positives also decrease on COCO and TinyPerson, while VisDrone incurs only one additional false positive. CQTR therefore improves small-object

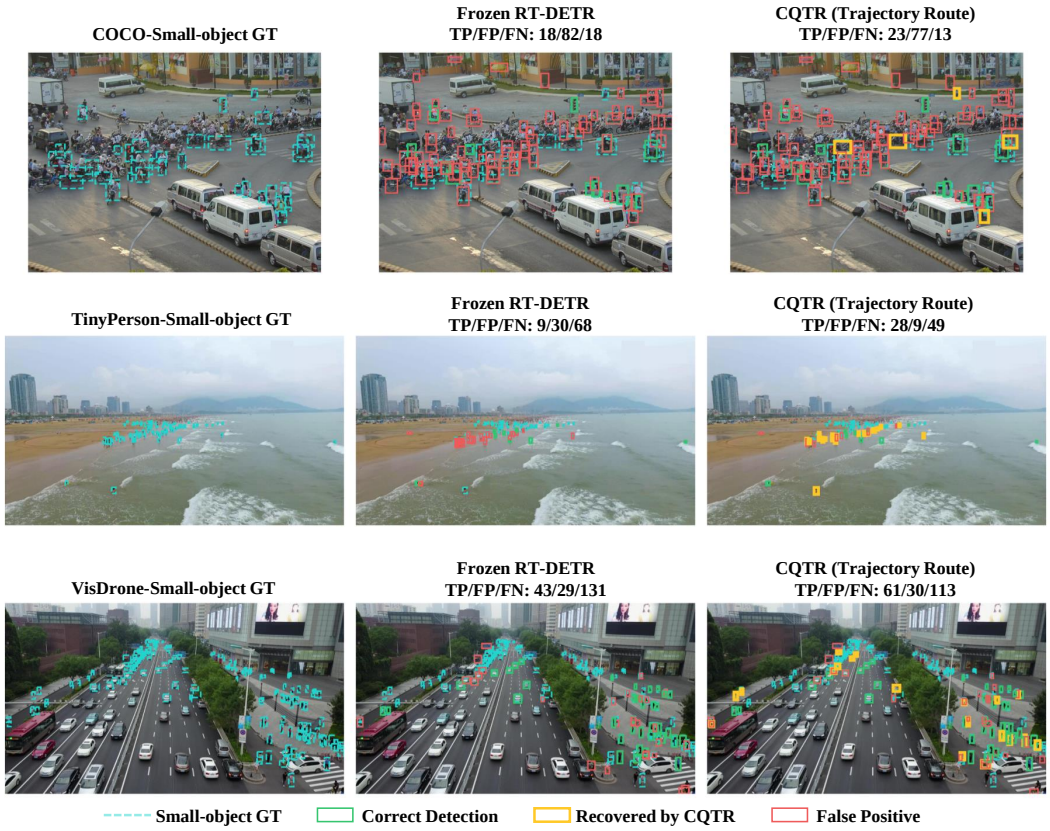

Figure 5: Final small-object detections produced by CQTR. Automatically selected examples at a shared confidence threshold show that CQTR recovers missed objects across all three datasets and reduces false positives on COCO and TinyPerson.

recall in crowded scenes and complex backgrounds while maintaining prediction reliability. The corrections motivated by internal interpretation ultimately translate into observable detection gains.

4.4 Mechanistic Interpretability and Analysis

Activation of latent scale knowledge. Fig. 6 compares RT-DETR-R50 under native inference and the counterfactual scale branch alone. Without updating any parameter, scale intervention improves both AP and AP_s on COCO, TinyPerson, and VisDrone. The AP_s gain is especially pronounced on COCO. These consistent improvements show that the relevant capability can be elicited from frozen parameters rather than necessarily relearned through target-domain training. The result supports our hypothesis that frozen detectors contain scale-sensitive representations that standard inference does not fully activate.

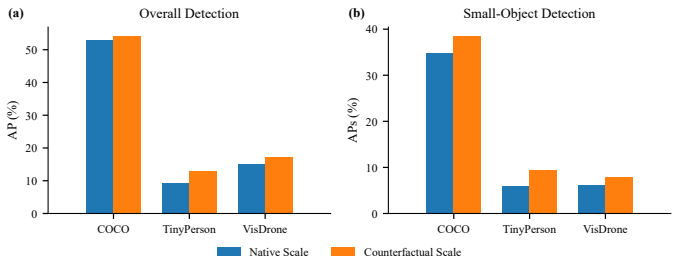

Figure 6: Activation of latent scale knowledge. AP and AP_s are compared between the frozen baseline and the counterfactual scale branch alone.

Reliability of query-trajectory evidence. We next test whether spatial-semantic trajectories explain candidate reliability. Using a fixed random seed, we sample 100 images from each evaluation set. Candidates with an area no greater than 32^2 and a confidence of at least 0.05 are divided into five equal-frequency bins according to trajectory evidence.

A candidate receives ground-truth support when its IoU with a small ground-truth object of the same class is at least 0.5. These annotations are used only for post hoc analysis and do not participate in CQTR inference or routing. As shown in Fig. 7, the small-object support rate generally rises with trajectory evidence. From the lowest to the highest evidence bin, the support rate increases from 1.35% to 17.62% on COCO, from 0.54% to 18.98% on TinyPerson, and from 4.82% to 30.84% on VisDrone. Trajectory evidence is therefore not an arbitrary score adjustment. It is directly associated with the correctness of small-object candidates.

Closed loop of unlabeled stream routing. Fig. 8 tests whether unlabeled stream statistics identify the more effective interpretation mechanism. The left panel reports normalized trajectory-evidence dispersion $\bar{d}/\tau_d$ and average maximum conflict strength $\bar{r}/\tau_c$, computed from an independent unlabeled training subset. The right panel shows the AP change of each correction branch relative to the scale-only branch. Both COCO statistics exceed their thresholds, so the router selects trajectory recalibration. TinyPerson and VisDrone are routed to conflict suppression. For these three RT-DETR-R50 streams, every routing decision agrees with the branch that performs better under full evaluation. Internal statistics derived from unlabeled images can therefore predict the effective correction mechanism, closing the loop from internal state to routing decision and performance improvement without selecting a branch post hoc from test AP.

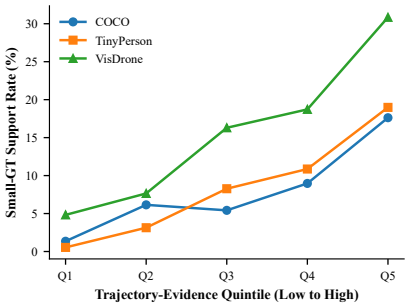

Figure 7: Query-trajectory evidence versus small-object ground-truth support. Q1 and Q5 denote the lowest and highest trajectory-evidence quintiles.

Decoder-internal evidence. We further visualize semantic query density and spatial trajectories across the final four decoder layers. High-evidence queries remain concentrated around small objects and converge more stably than low-evidence queries, providing direct internal support for trajectory-based reliability estimation. Detailed definitions, statistics, and visualizations are provided in Appendix A.2.

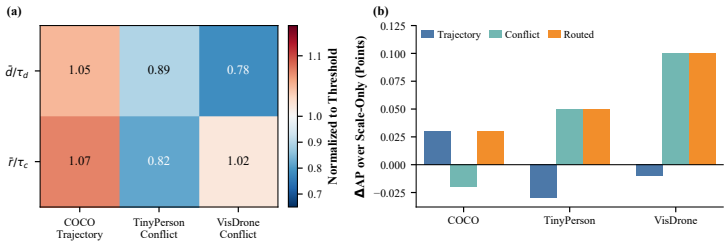

Figure 8: Closed loop of unlabeled stream routing. (a) Routing statistics normalized by their thresholds. (b) AP changes of the trajectory, conflict, and routed branches relative to the scale-only branch.

5 Conclusion

We presented CQTR, a training-free framework for improving small-object detection with frozen detectors. Through a counterfactual scale branch, spatial-semantic query-trajectory interpretation, and unlabeled reliability routing, CQTR activates latent scale knowledge and suppresses cross-scale conflicts. Experiments across multiple detectors and three datasets demonstrate its effectiveness, generality, and interpretability. Future work will investigate lower-cost single-pass interventions and extensions to open-world detection.

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

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*, pp. 966–970, 2022.

Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. SOD-MTGAN: Small object detection via multi-task generative adversarial network. In *Proceedings of the European Conference on Computer Vision*, pp. 206–221, 2018.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6541–6549, 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*, pp. 213–229, 2020.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bimodal and encoder-decoder transformers. In *IEEE/CVF International Conference on Computer Vision*, pp. 397–406, 2021a.

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, 2021b.

Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic DETR: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2988–2997, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, Liefeng Bo, Hailin Shi, Rui Zhu, Aashish Kumar, Ajin Li, Almaz Zillolayev, Anuar Askergaliev, Arne Schumann, Binjie Mao, Byeongwon Lee, Chang Liu, Changrui Chen, Chunhong Pan, Chunlei Huo, Da Yu, Dechun Cong, Dening Zeng, Dheeraj Reddy Pailla, Di Li, Dong Wang, Donghyeon Cho, Dongyu Zhang, Furui Bai, George Jose, Guangyu Gao, Guizhong Liu, Haitao Xiong, Hao Qi, Haoran Wang, Heqian Qiu, Hongliang Li, Huchuan Lu, Ildoo Kim, Jaekyum Kim, Jane Shen, Jihoon Lee, Jing Ge, Jingjing Xu, Jingkai Zhou, Jonas Meier, Jun Won Choi, Junhao Hu, Junyi Zhang, Junying Huang, Kaiqi Huang, Keyang Wang, Lars Sommer, Lei Jin, Lei Zhang, Lianghua Huang, Lin Sun, Lucas Steinmann, Meixia Jia, Nuo Xu, Pengyi Zhang, Qiang Chen, Qingxuan Lv, Qiong Liu, Qishang Cheng, Sai Saketh Chennamsetty, Shuhao Chen, Shuo Wei, Srinivas S. S. Kruthiventi, Sungeun Hong, Sungil Kang, Tong Wu, Tuo Feng, Varghese Alex Kollerathu, Wanqi Li, Wei Dai, Weida Qin, Weiyang Wang, Xiaorui Wang, Xiaoyu Chen, Xin Chen, Xin Sun, Xin Zhang, Xin Zhao, Xindi Zhang, Xinyu Zhang, Xuankun Chen, Xudong Wei, Xuzhang Zhang, Yanchao Li, Yifu Chen, Yu Heng Toh, Yu Zhang, Yu Zhu, Yunxin Zhong, Zexin Wang, Zhikang Wang, Zichen Song, and Ziming Liu. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In *IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

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.

Xiuquan Hou, Meiqin Liu, Senlin Zhang, Ping Wei, and Badong Chen. Salience DETR: Enhancing detection transformer with hierarchical salience filtering refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17574–17583, 2024.

Shihua Huang, Zhichao Lu, Xiaodong Cun, Yongjun Yu, Xiao Zhou, and Xi Shen. DEIM: DETR with improved matching for fast convergence. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15162–15171, 2025.

Ashkan Khakzar, Soroosh Baselizadeh, Saurabh Khanduja, Christian Rupprecht, Seong Tae Kim, and Nassir Navab. Neural response interpretation through the lens of critical pathways. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13528–13538, 2021.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677, 2018.

Mate Kisantal, Zbigniew Wojna, Jakub Murawski, Jacek Naruniec, and Kyunghyun Cho. Augmentation for small object detection. *arXiv preprint arXiv:1902.07296*, 2019.

Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, and Lei Zhang. DN-DETR: Accelerate DETR training by introducing query denoising. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13619–13627, 2022.

Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Perceptual generative adversarial networks for small object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1222–1230, 2017.

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, 2014.

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, 2017.

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.

Wenyu Lv, Yian Zhao, Qinyao Chang, Kui Huang, Guanzhong Wang, and Yi Liu. RT-DETRv2: Improved baseline with bag-of-freebies for real-time detection transformer. *arXiv preprint arXiv:2407.17140*, 2024.

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*, pp. 3651–3660, 2021.

Mahyar Najibi, Bharat Singh, and Larry S. Davis. AutoFocus: Efficient multi-scale inference. In *IEEE/CVF International Conference on Computer Vision*, pp. 9745–9755, 2019.

Yansong Peng, Hebei Li, Peixi Wu, Yueyi Zhang, Xiaoyan Sun, and Feng Wu. D-FINE: Redefine regression task of DETRs as fine-grained distribution refinement. In *International Conference on Learning Representations*, 2025.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems*, volume 34, pp. 12116–12128, 2021.

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*, pp. 618–626, 2017.

Tharsan Senthivel and Ngoc-Son Vu. QR-DETR: Query routing for detection transformer. In *Proceedings of the Asian Conference on Computer Vision*, pp. 354–371, 2024.

Bharat Singh and Larry S. Davis. An analysis of scale invariance in object detection—SNIP. In *IEEE/CVF 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, 2018.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, 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.

Shuo Wang, Chunlong Xia, Feng Lv, and Yifeng Shi. RT-DETRv3: Real-time end-to-end object detection with hierarchical dense positive supervision. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1628–1636, 2025.

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

Fan Yang, Heng Fan, Peng Chu, Erik Blasch, and Haibin Ling. Clustered object detection in aerial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8311–8320, 2019.

Mingqiao Ye, Lei Ke, Siyuan Li, Yu-Wing Tai, Chi-Keung Tang, Martin Danelljan, and Fisher Yu. Cascade-DETR: Delving into high-quality universal object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6704–6714, 2023.

Xuehui Yu, Yuqi Gong, Nan Jiang, Qixiang Ye, and Zhenjun Han. Scale match for tiny person detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1257–1265, 2020.

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.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. DETRs beat YOLOs on real-time object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16965–16974, 2024.

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

A.1 Full Quantitative Results

Figures 3 and 4 in the main paper visualize the overall AP and small-object AP trends across the three datasets. To provide complete quantitative evidence, Tables 2–4 report AP, AP_{50}, AP_{75}, and AP_s for all 27 detector-dataset combinations. Values in parentheses denote absolute improvements over the corresponding frozen baselines. All results are obtained without updating the model parameters or using target-domain training annotations.

COCO results. Table 2 provides the complete results on COCO val2017. CQTR improves all four evaluation metrics for every frozen detector. The highest AP of 54.32 is obtained by RT-DETRv2-R50 with CQTR. RT-DETR-R50 exhibits the largest $\mathrm{AP_s}$ improvement, increasing from 34.74 to 38.55. These detailed results confirm the trend in Figures 3 and and 4, where the improvement in small-object detection is more pronounced than the overall AP gain.

Cross-domain TinyPerson results. Table 3 reports the complete cross-domain results on TinyPerson. CQTR consistently improves AP, AP_{50}, AP_{75}, and $\mathrm{AP_s}$ across all nine detectors. RT-DETR-R50 achieves the largest improvement, with AP increasing from 9.19 to 13.04 and $\mathrm{AP_s}$ increasing from 5.89 to 9.60. The consistent gains across different detector families demonstrate that CQTR transfers scale-sensitive knowledge beyond the COCO pretraining distribution.

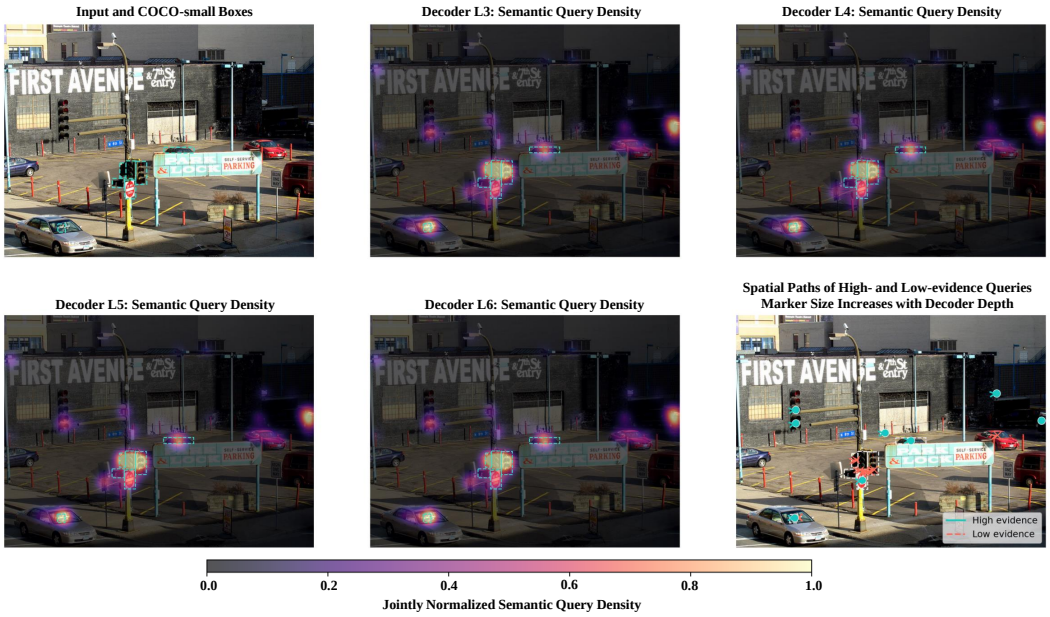

Figure 9: Decoder-internal spatial-semantic query trajectories. Semantic-density maps and representative trajectories from the final four decoder layers show that high-evidence queries converge more stably around small objects than low-evidence queries, with ground truth used only for post hoc visualization.

Cross-domain VisDrone2019 results. As reported in Table 4, CQTR also improves every detector on VisDrone2019-DET val. DEIM-D-FINE-S obtains the largest AP gain, increasing from 14.50 to 17.16, together with an $\mathrm{AP_s}$ improvement from 6.22 to 8.61. RT-DETRv2-R50 achieves the highest absolute AP of 17.49. These results support the effectiveness of CQTR in dense aerial scenes with severe scale variation and background interference.

A.2 Visualization of Decoder Query Trajectories

To further explain the internal meaning of spatial-semantic trajectory evidence, we project query states from the final four decoder layers onto the original image. Let $\mu_q^{(l)}$ be the center of the box predicted by query q at layer l, and let $p_q^{(l)}(y_q)$ denote the response at that layer to its final class y_q. The semantic query density at layer l is defined as

$$H^{(l)}(u,v) = \sum_{q \in Q_s} p_q^{(l)}(y_q) \exp\left(-\frac{\|(u,v) - \mu_q^{(l)}\|_2^2}{2\sigma^2} \right), \tag{9}$$

where Q_s is the set of queries whose final predicted area does not exceed the small-object threshold. All layers use a shared normalization range, allowing their colors to be compared directly.

As shown in Fig. 9, decoder query responses concentrate on small-object regions such as traffic lights and vehicles, and maintain high object-to-background response ratios in later layers. The image contains 80 small-object candidate queries. Their response ratios between small-object and background regions are 24.58, 26.44, 26.14, and 26.06 across the four layers, showing persistent concentration around small objects during iterative decoding. The lower-right panel further compares the spatial paths of high- and low-evidence queries. High-evidence queries remain spatially convergent across the last four layers, whereas low-evidence queries move back and forth more noticeably in the dense traffic-light region. The mean trajectory length of high-evidence queries is only 0.013% of the image diagonal. The corresponding value for low-evidence queries is 0.184%, approximately 14.3 times larger. CQTR trajectory evidence thus separates stably convergent queries from spatially uncertain ones and provides observable internal support for subsequent reliability recalibration.

Table 2: Comparison on COCO val2017.

Method	AP	AP_{50}	AP_{75}	AP_s
DETR-R50	36.51	55.83	37.74	13.65
DETR-R50 + CQTR	38.16 (+1.65)	57.51 (+1.68)	40.17 (+2.43)	15.77 (+2.12)
Deformable-DETR-R50	40.45	58.58	43.83	19.44
Deformable-DETR-R50 + CQTR	41.98 (+1.53)	60.26 (+1.68)	45.78 (+1.95)	22.08 (+2.64)
DAB-DETR-R50	36.35	56.40	37.64	13.61
DAB-DETR-R50 + CQTR	37.94 (+1.59)	57.72 (+1.32)	39.97 (+2.33)	15.74 (+2.13)
DINO-R50	44.62	61.70	48.26	24.03
DINO-R50 + CQTR	46.23 (+1.61)	63.32 (+1.62)	50.33 (+2.07)	26.51 (+2.48)
RT-DETR-R50	53.09	71.23	57.73	34.74
RT-DETR-R50 + CQTR	54.21 (+1.12)	71.86 (+0.63)	59.19 (+1.46)	38.55 (+3.81)
RT-DETRv2-R50	53.40	71.56	57.48	36.06
RT-DETRv2-R50 + CQTR	54.32 (+0.92)	72.22 (+0.66)	58.88 (+1.40)	38.01 (+1.95)
RT-DETRv3-R50	52.72	70.95	57.26	35.15
RT-DETRv3-R50 + CQTR	53.72 (+1.00)	71.55 (+0.60)	58.68 (+1.42)	38.15 (+3.00)
D-FINE-S	48.49	65.42	52.58	29.30
D-FINE-S + CQTR	49.58 (+1.09)	66.48 (+1.06)	53.94 (+1.36)	31.70 (+2.40)
DEIM-D-FINE-S	48.96	65.88	53.04	30.38
DEIM-D-FINE-S + CQTR	49.71 (+0.75)	66.69 (+0.81)	54.23 (+1.19)	32.57 (+2.19)

Table 3: Cross-domain COCO-style comparison on TinyPerson test.

Method	AP	AP_{50}	AP_{75}	AP_s
DETR-R50	1.87	5.39	1.12	0.69
DETR-R50 + CQTR	2.72 (+0.85)	8.02 (+2.63)	1.22 (+0.10)	1.26 (+0.57)
Deformable-DETR-R50	2.92	9.22	1.13	1.09
Deformable-DETR-R50 + CQTR	4.45 (+1.53)	13.37 (+4.15)	1.89 (+0.76)	2.02 (+0.93)
DAB-DETR-R50	1.78	5.34	1.11	0.67
DAB-DETR-R50 + CQTR	3.01 (+1.23)	9.72 (+4.38)	1.35 (+0.24)	1.11 (+0.44)
DINO-R50	4.28	13.56	1.98	2.11
DINO-R50 + CQTR	6.71 (+2.43)	20.50 (+6.94)	3.05 (+1.07)	3.89 (+1.78)
RT-DETR-R50	9.19	26.24	4.63	5.89
RT-DETR-R50 + CQTR	13.04 (+3.85)	34.83 (+8.59)	7.11 (+2.48)	9.60 (+3.71)
RT-DETRv2-R50	9.45	26.16	4.82	5.97
RT-DETRv2-R50 + CQTR	11.66 (+2.21)	32.23 (+6.07)	5.96 (+1.14)	7.88 (+1.91)
RT-DETRv3-R50	9.12	25.37	4.74	5.80
RT-DETRv3-R50 + CQTR	12.82 (+3.70)	33.98 (+8.61)	7.28 (+2.54)	9.15 (+3.35)
D-FINE-S	7.24	20.74	3.67	4.37
D-FINE-S + CQTR	9.80 (+2.56)	27.39 (+6.65)	4.99 (+1.32)	6.77 (+2.40)
DEIM-D-FINE-S	7.80	21.88	4.15	4.96
DEIM-D-FINE-S + CQTR	9.80 (+2.00)	27.29 (+5.41)	5.39 (+1.24)	6.73 (+1.77)

Table 4: Cross-domain COCO-style comparison on VisDrone2019-DET val.

Method	AP	AP_{50}	AP_{75}	AP_s
DETR-R50	4.31	8.95	3.65	0.60
DETR-R50 + CQTR	5.12 (+0.81)	10.86 (+1.91)	4.20 (+0.55)	0.96 (+0.36)
Deformable-DETR-R50	5.74	11.75	4.88	1.34
Deformable-DETR-R50 + CQTR	6.99 (+1.25)	14.17 (+2.42)	6.13 (+1.25)	2.04 (+0.70)
DAB-DETR-R50	4.59	9.54	4.02	0.68
DAB-DETR-R50 + CQTR	6.14 (+1.55)	12.73 (+3.19)	5.23 (+1.21)	1.11 (+0.43)
DINO-R50	7.86	15.53	7.03	2.36
DINO-R50 + CQTR	9.25 (+1.39)	18.14 (+2.61)	8.33 (+1.30)	3.65 (+1.29)
RT-DETR-R50	15.11	27.11	14.57	6.16
RT-DETR-R50 + CQTR	17.23 (+2.12)	30.80 (+3.69)	16.46 (+1.89)	7.85 (+1.69)
RT-DETRv2-R50	15.06	27.16	14.43	6.02
RT-DETRv2-R50 + CQTR	17.49 (+2.43)	30.95 (+3.79)	16.85 (+2.42)	8.10 (+2.08)
RT-DETRv3-R50	15.07	26.83	14.59	6.22
RT-DETRv3-R50 + CQTR	17.27 (+2.20)	30.81 (+3.98)	16.68 (+2.09)	7.92 (+1.70)
D-FINE-S	12.98	23.55	12.44	5.55
D-FINE-S + CQTR	15.27 (+2.29)	26.90 (+3.35)	14.87 (+2.43)	7.53 (+1.98)
DEIM-D-FINE-S	14.50	26.24	13.68	6.22
DEIM-D-FINE-S + CQTR	17.16 (+2.66)	30.05 (+3.81)	16.84 (+3.16)	8.61 (+2.39)