A GAZE-FREE HUMAN-INSPIRED OBJECT DETECTION EXPLAINER VIA EVIDENCE-BASED FIXATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Explanations for object detectors are usually evaluated against detector behavior, leaving their alignment with human attention unclear. Human-inspired explainers largely target image classification, while human-supervised methods learn from gaze and cannot use it as an independent reference. We introduce HiOdex *(Human-inspired Object Detection Explainer)*, the first post-hoc, human-inspired explainer that produces object-specific attributions for object detectors without gaze supervision. HiOdex derives model-generated fixations from activation-gradient evidence, applies inhibition of return, and uses their envelope to guide class activation mapping (CAM). Local evidence and multi-layer fusion preserve spatial detail in a single forward and backward pass. We compare HiOdex with eight explainers across two detectors and two benchmarks using model-oriented metrics, then independently assess human-gaze alignment on two gaze datasets using image-level saliency metrics and object-level measures. HiOdex achieves the highest Over-All score among all evaluated CAM-based and human-inspired baselines in all four detector and benchmark settings. It comes within 0.0217 of the strongest perturbation-based result at roughly 1% of the computational cost, achieves the best or tied-best AUC-Judd in every gaze setting, and ranks in the top two for 8 of 12 evaluations using image-level gaze metrics. These results show detector-derived fixations can yield efficient explanations that perform strongly under both model- and human-centered evaluation without gaze supervision.

1 INTRODUCTION

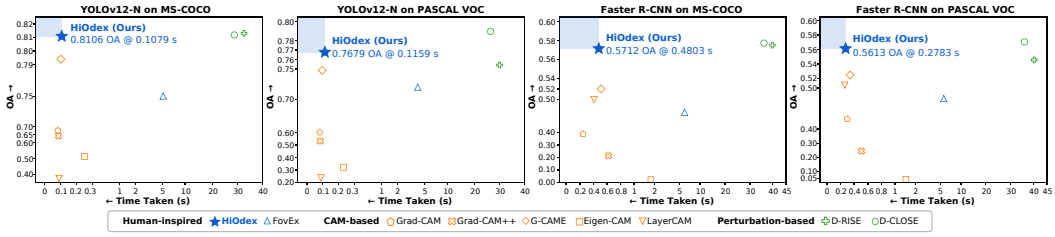

Figure 1: **HiOdex gives the best explanation quality-cost trade-off across all four settings.** It attains the highest Over-All (OA) score among all evaluated CAM-based and human-inspired baselines, and no evaluated method achieves both a lower generation time and a higher OA. OA is plotted against mean attribution-generation time per explanation for both detectors and benchmarks, with better trade-offs toward the upper-left corner. Both axes use compressed non-uniform scales.

Visual explanations for object detectors ask which image regions support a selected prediction. Model-oriented metrics (Bommer et al., 2024; Nguyen et al., 2025) quantify this relation, but they do not answer its human-centered counterpart: *"Does the explanation highlight the regions a person would inspect (Kazmierczak et al., 2026)?"* This distinction is especially important in object detection, where each explanation must correspond to one category-and-location prediction among potentially many objects. Gradient-based methods such as Grad-CAM (Selvaraju et al., 2017), G-CAME (Nguyen et al., 2024a), and ODAM (Zhao et al., 2024) obtain an attribution from internal

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.

activations and gradients, often with one forward and one backward pass. Perturbation-based methods such as D-RISE (Petsiuk et al., 2021) and D-CLOSE (Nguyen et al., 2024b) instead estimate importance from prediction changes across hundreds or thousands of masked inputs. Despite their different mechanisms and costs, both families are evaluated primarily against the detector itself.

Human gaze provides an independently collected behavioral reference for the complementary question. Eye fixations record where observers allocate attention (Judd et al., 2009), while established saliency metrics (Bylinskii et al., 2018) compare fixation patterns with attribution maps. Existing human-oriented explanation methods follow two distinct directions. Human-inspired methods incorporate mechanisms modeled on visual behavior. FovEx (Panda et al., 2025), for example, selects gradient-informed fixations through iterative image foveation, but primarily addresses image classification and does not evaluate gaze alignment in its limited object-detection study. Human-supervised methods instead learn from recorded attention. HAG-XAI (Liu et al., 2024) optimizes saliency toward human-attention maps, while HINT (Selvaraju et al., 2019) aligns gradient importance with human annotations during training. The attention data used to train or tune these methods cannot subsequently serve as an independent evaluation reference. To our knowledge, no post-hoc, annotation-free method embeds fixation behavior directly into attribution for an individual detection, and no detector explainer has been evaluated against human gaze at the dataset scale. This gap leaves unresolved whether efficient detector explanations can be both model-grounded and human-aligned.

We propose HiOdex *(Human-inspired Object Detection Explainer)*, which produces an object-specific attribution map for a selected prediction from a differentiable detector. HiOdex places its human-inspired mechanism within the attribution process rather than applying foveation to a completed saliency map. Positive activation-gradient evidence drives a sequence of model-generated fixations, while inhibition of return distributes them across distinct evidence-bearing regions. Their spatial envelope guides gradient pooling into class activation mapping (CAM) channel weights. A complementary local-evidence branch preserves position-wise structure before object gating and multi-layer fusion. HiOdex uses fixed hyperparameters across detectors, datasets, and instances, requires only one forward and one backward pass, and never uses recorded gaze. Human fixations therefore remain independent evidence for evaluating the resulting explanations.

Our contributions are summarized as follows: (1) **Object-specific human-inspired attribution:** We introduce a post-hoc explainer that combines evidence-driven fixation selection, inhibition of return, and fixation-envelope-guided gradient pooling without gaze supervision. (2) **Cross-architecture evaluation:** We compare HiOdex with eight representative gradient-based, perturbation-based, and human-inspired explainers using the one-stage YOLOv12-N (Tian et al., 2026) and two-stage Faster R-CNN (Ren et al., 2016) on MS-COCO (Lin et al., 2014) and PASCAL VOC (Everingham et al., 2010). (3) **Independent gaze evaluation:** We conduct dataset-scale evaluation on COCO-Search18 (Chen et al., 2021) and POET (Papadopoulos et al., 2014) using established image- and object-level saliency metrics, together with two proposed object-level measures, Object-Focused Cluster Score (OFCS) and Weighted Distribution Similarity (WDSIM). (4) **Empirical performance and efficiency:** HiOdex attains the highest Over-All (OA) score (Zhang et al., 2021) among all evaluated CAM-based methods and FovEx in every detector and benchmark setting, remains within 0.0217 of the strongest perturbation-based result at approximately 1% of the computational cost (Fig. 1), and ranks among the top three in 22 of 24 model-oriented comparisons. It also achieves the best or tied-best AUC-Judd (Judd et al., 2009) in all four gaze settings and ranks among the top two in 8 of 12 comparisons using image-level gaze metrics.

2 Background and Related Work

2.1 Visual XAI for Object Detection Models

Visual explanations for object detection models primarily rely on gradient- or perturbation-based strategies. Grad-CAM (Selvaraju et al., 2017) globally pools gradients into CAM channel weights. Grad-CAM++ (Chattopadhay et al., 2018) uses location-dependent gradient weighting, LayerCAM (Jiang et al., 2021) preserves position-wise positive gradients, and Eigen-CAM (Muhammad & Yeasin, 2020) extracts the principal activation component without class-specific gradients. Because these methods were developed primarily for image classification, their explanations identify class evidence but do not inherently distinguish individual detections of the same category.

Detection-specific methods condition attribution on a selected prediction. G-CAME (Nguyen et al., 2024a) applies a target-centered Gaussian to activation-derived saliency, while ODAM (Zhao et al., 2024) backpropagates an instance-specific detection target into intermediate features. Both retain the low cost of gradient-based attribution. Perturbation-based methods instead infer importance from changes in detector output. D-RISE (Petsiuk et al., 2021) weights random masks by their agreement with the target detection, and D-CLOSE (Nguyen et al., 2024b) aggregates evidence across multiple segmentation granularities. These approaches can provide strong faithfulness, but require hundreds or thousands of detector evaluations per explanation rather than one forward and one backward pass (Petsiuk et al., 2021; Nguyen et al., 2024b).

Explanation Metrics in Object Detection Model-oriented evaluation captures complementary properties of an explanation. Insertion and Deletion (Petsiuk et al., 2018) measure how the target prediction changes as highly attributed regions are revealed or removed, and Over-All (OA) (Zhang et al., 2021) summarizes the difference between their areas under the curve. The Pointing Game (PG) (Zhang et al., 2018) tests whether the attribution maximum lies inside the target object, while its energy-based variant (EBPG) (Wang et al., 2020) measures the attribution mass within that region. Sparsity (Gomez et al., 2022) quantifies attribution concentration. No method consistently dominates these criteria (Bommer et al., 2024; Nguyen et al., 2025), and all remain model-oriented because they compare an explanation with detector behavior or object annotations rather than with human visual attention (Kazmierczak et al., 2026).

Object Detection Architectures Object detectors expose prediction-specific evidence at different stages of their feature hierarchies. One-stage detectors such as YOLO (Redmon et al., 2016) predict categories and locations directly from dense, multi-scale features, whereas two-stage detectors such as Faster R-CNN (Ren et al., 2016) classify and refine region proposals using features pooled from a shared backbone. Our evaluation represents these paradigms with YOLOv12-N (Tian et al., 2026), whose attention-centric design retains real-time inference, and Faster R-CNN with a Feature Pyramid Network (FPN). Their distinct prediction pipelines test whether an explainer can associate internal evidence with one selected detection rather than only with its category.

2.2 Human-Inspired XAI in Object Detection

Human gaze provides an independently collected behavioral reference for whether an explanation emphasizes regions that people inspect. AUC-Judd (Judd et al., 2009) evaluates whether fixated locations receive greater attribution than non-fixated locations, Similarity (SIM) (Bylinskii et al., 2018) measures overlap between normalized spatial distributions, and Kullback-Leibler (KL) divergence measures their discrepancy. These metrics capture different properties and should be interpreted jointly (Bylinskii et al., 2018). However, they were developed for whole-image saliency and do not identify the object associated with a particular detection. We therefore use them as the primary gaze-alignment criteria and complement them with object-level measures restricted to the selected detection.

Human visual behavior enters explanation methods in two distinct ways. Human-inspired methods reproduce selected perceptual mechanisms without learning from recorded attention. FovEx (Panda et al., 2025) selects gradient-informed fixations, iteratively foveates the input, and aggregates the resulting explanations. Its evaluation primarily concerns image classification, while its limited detection study neither performs dataset-scale evaluation nor measures gaze alignment. Human-supervised methods instead learn from attention annotations. HAG-XAI (Liu et al., 2024) optimizes saliency toward human-attention maps for classification and detection, whereas HINT (Selvaraju et al., 2019)aligns gradient importance with human annotations during training for vision-language tasks. Such supervision can improve human alignment, but the attention data used for optimization can no longer provide an independent evaluation reference.

Positioning of HiOdex HiOdex is the first post-hoc, human-inspired explainer that produces object-specific attributions for object detection without gaze supervision. Existing detector explainers produce object-specific attribution but do not model how attention moves across distinct evidence-bearing regions of the selected object. FovEx (Panda et al., 2025) produces model-generated fixations, but uses them to transform the input before aggregating whole-image explanations. HiOdex instead embeds evidence-driven fixation selection and inhibition of return within

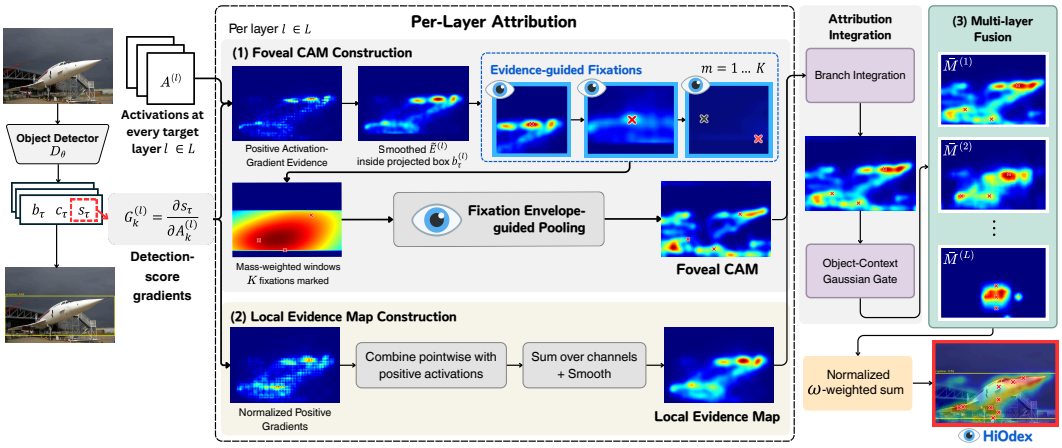

Figure 2: **Overview of HiOdex.** For a selected detection $\tau = (b_\tau, c_\tau, s_\tau)$, one forward and backward pass supplies activations and detection-score gradients at every target layer. The Foveal CAM converts positive activation-gradient evidence into model-generated fixations and an envelope for spatial gradient pooling, while the Local Evidence Map retains position-wise structure. Branch integration, object-context gating, and multi-layer fusion produce the final heatmap. The pooling support is therefore determined by detector evidence rather than recorded human gaze. Crosses mark model-generated fixations.

detector-feature attribution. Its fixation envelope controls where gradients are pooled into CAM channel weights, rather than reweighting a completed map as in G-CAME (Nguyen et al., 2024a), while a local-evidence branch preserves position-wise structure. HiOdex therefore remains object-specific and uses no recorded gaze, preserving human fixations as independent evaluation evidence.

3 Method

We next describe the formulation and main components of HiOdex. Fig. 2 summarizes the overall pipeline. For a selected detection, HiOdex constructs a fixation-guided Foveal CAM and a Local Evidence Map, integrates them at each feature level, and applies object-context refinement and multi-layer fusion to obtain the final attribution. We detail these stages below.

3.1 Preliminaries

Let $D_\theta : \mathcal{X} \to \mathcal{Y}$ be a differentiable detector and $x \in \mathbb{R}^{H \times W \times 3}$ an input image. We explain one selected prediction $\tau = (b_\tau, c_\tau, s_\tau)$, comprising a bounding box $b_\tau = (x_1, y_1, x_2, y_2)$, predicted class c_τ, and differentiable score $s_\tau = \psi_\tau(D_\theta(x))$ before non-maximum suppression. The function ψ_τ extracts the corresponding logit, class score, or confidence. HiOdex only requires s_τ to be differentiable with respect to the selected spatial feature layers.

For each target layer $l \in \mathcal{L}$, let $A^{(l)} \in \mathbb{R}^{C_l \times H_l \times W_l}$ be its activation tensor and $G_k^{(l)} = \partial s_\tau / \partial A_k^{(l)}$ the gradient for channel k. A single forward and backward pass yields $A^{(l)}$ and $G^{(l)}$ for every $l \in \mathcal{L}$. We denote spatial min-max normalization to $[0, 1]$ by $\mathrm{Norm}(\cdot)$, isotropic Gaussian smoothing by $\mathrm{GaussianBlur}(\cdot, \sigma)$, element-wise multiplication by $\odot$, and a small numerical constant by ε. Standard deviations are measured in feature cells for layer maps and pixels for image-space maps.

3.2 Foveal CAM Construction

The Foveal CAM replaces global spatial averaging of gradients with evidence-conditioned pooling. It first generates fixations from the detector's own evidence, then pools gradients within their spatial envelope.

Object evidence At layer l, the positive activation-gradient evidence is:

$$E^{(l)}(i,j) = \sum_{k=1}^{C_l} \mathrm{ReLU}\Big(G_k^{(l)}(i,j)\Big)\,\mathrm{ReLU}\Big(A_k^{(l)}(i,j)\Big), \quad \widetilde{E}^{(l)} = \mathrm{GaussianBlur}\Big(E^{(l)}, 1\Big). \quad (1)$$

Smoothing prevents isolated feature cells from dominating fixation selection. We project b_τ onto the layer grid as $b_\tau^{(l)} = (x_1^{(l)}, y_1^{(l)}, x_2^{(l)}, y_2^{(l)})$, with width $w_l = x_2^{(l)} - x_1^{(l)}$ and height $h_l = y_2^{(l)} - y_1^{(l)}$.

Evidence-guided fixations HiOdex selects K fixations from the smoothed evidence inside $b_\tau^{(l)}$. The initial working map $W_1^{(l)}$ equals $\widetilde{E}^{(l)}$ inside the projected box and zero elsewhere. At step m, HiOdex selects the strongest remaining location and constructs an object-scaled anisotropic window:

$$f_m^{(l)} = (x_m, y_m) = \arg\max_{(i,j)} W_m^{(l)}(i,j), \quad (2)$$

$$Q_m^{(l)}(i,j) = \exp\left[-\frac{(j-x_m)^2}{2(\sigma_x^{(l)})^2} - \frac{(i-y_m)^2}{2(\sigma_y^{(l)})^2}\right], \quad \sigma_x^{(l)} = \delta w_l, \quad \sigma_y^{(l)} = \delta h_l. \quad (3)$$

The window follows the selected object's extent rather than the full image. Inhibition of return updates the evidence as $W_{m+1}^{(l)} = W_m^{(l)} \odot (1 - Q_m^{(l)})$, encouraging successive fixations to cover distinct evidence-bearing regions.

Fixation envelope Each fixation is weighted by the evidence mass captured by its window. With $m_m^{(l)} = \sum_{i,j} Q_m^{(l)}(i,j) W_1^{(l)}(i,j)$ and $\pi_m^{(l)} = m_m^{(l)} / (\sum_{n=1}^{K} m_n^{(l)} + \varepsilon)$, the envelope is:

$$Q^{(l)}(i,j) = \begin{cases} \sum_{m=1}^{K} \pi_m^{(l)} Q_m^{(l)}(i,j), & (i,j) \in b_\tau^{(l)}, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

The envelope therefore distributes support across evidence-bearing regions while remaining confined to the selected detection.

Fixation-envelope-guided pooling The defining operation of HiOdex pools gradients within this envelope:

$$\begin{aligned} \alpha_k^{(l)} &= \frac{\sum_{i,j} Q^{(l)}(i,j) G_k^{(l)}(i,j)}{\sum_{i,j} Q^{(l)}(i,j) + \varepsilon}, \\ H_F^{(l)} &= \mathrm{GaussianBlur}\left(\mathrm{ReLU}\left(\sum_{k=1}^{C_l} \alpha_k^{(l)} A_k^{(l)}\right), \gamma \min(H_l, W_l)\right). \end{aligned} \quad (5)$$

Compared with global gradient pooling, this operation gives greater influence to gradients at evidence-supported locations. The denominator makes the channel weights invariant to the overall scale of $Q^{(l)}$.

3.3 Local Evidence Map Construction

The Foveal CAM is detection-specific, but channel aggregation can suppress fine spatial variation. The Local Evidence Map retains this variation by normalizing positive gradients within each channel and combining them pointwise with positive activations:

$$\begin{aligned} \widehat{G}_k^{(l)}(i,j) &= \frac{\mathrm{ReLU}(G_k^{(l)}(i,j))}{\max_{a,b} \mathrm{ReLU}(G_k^{(l)}(a,b)) + \varepsilon}, \\ H_L^{(l)} &= \mathrm{GaussianBlur}\left(\sum_{k=1}^{C_l} \widehat{G}_k^{(l)} \odot \mathrm{ReLU}(A_k^{(l)}), \gamma \min(H_l, W_l)\right). \end{aligned} \quad (6)$$

Smoothing suppresses fragmented cell-level responses while preserving position-wise evidence that channel-level aggregation may lose.

3.4 Attribution Integration and Multi-Layer Fusion

At each target layer, HiOdex integrates the normalized branches, resizes the result bilinearly to the input resolution, and applies an object-context gate. For the center (c_x, c_y), width w_τ, and height h_τ of b_τ,

$$\begin{aligned} M^{(l)} &= \mathrm{Norm}\Big(\beta\,\mathrm{Norm}(H_F^{(l)}) + (1-\beta)\,\mathrm{Norm}(H_L^{(l)})\Big), \\ R(u,v) &= \exp\left[-\frac{(u-c_x)^2}{2(\rho w_\tau)^2} - \frac{(v-c_y)^2}{2(\rho h_\tau)^2}\right], \\ \bar{M}^{(l)} &= \mathrm{Norm}\Big(\widetilde{M}^{(l)} \odot R\Big), \qquad H = \mathrm{Norm}\left(\sum_{l\in\mathcal{L}} \omega_l \bar{M}^{(l)}\right). \end{aligned} \tag{7}$$

Here, β balances fixation-envelope-guided channel attribution and local evidence, while $\widetilde{M}^{(l)}$ denotes the resized map. The gate suppresses distant responses while retaining nearby context. The feature-resolution ordering determines ω_l, and the same rule is used across detectors, datasets, images, and objects. All operations after backpropagation act on the cached activations and gradients, so HiOdex requires no further detector evaluations.

4 Experimental Setup

Our experiments address (Q1) whether HiOdex provides competitive model-oriented explanation quality across detector architectures and datasets; (Q2) whether its explanations align with independently collected human gaze without gaze supervision; (Q3) whether this performance is achieved at practical computational cost; and (Q4) which components account for the observed behavior. Sec. 5 presents the empirical comparisons to address Q1-3, while App. C.3 presents the mechanism ablations that address Q4. App. B and C provide complete protocol and extended analysis.

Detectors and Datasets We use YOLOv12-N (Tian et al., 2026) and Faster R-CNN (Ren et al., 2016) with a ResNet-50 FPN backbone on the MS-COCO (Lin et al., 2014) and PASCAL VOC (Everingham et al., 2010) datasets. Every method is evaluated on the same fixed 1,000-image subset within each setting. Human-gaze alignment is assessed on the corresponding image domains using all eligible images from COCO-Search18 (Chen et al., 2021), which contains 6,202 MS-COCO images and observations from ten observers, and POET (Papadopoulos et al., 2014), which contains 6,270 PASCAL VOC images and observations from five observers. App. D.1 provides an additional evaluation on the Substation Equipment dataset (Gomes et al., 2023), containing many small, visually similar objects embedded in complex structures. Both detectors use official COCO-pretrained checkpoints and process 640×640 letterboxed inputs. Predictions are retained at a confidence threshold of 0.25 and processed using non-maximum suppression with an intersection-over-union threshold of 0.45. All CAM-based methods access the same detector-specific target layers, preventing layer selection from confounding their comparison.

XAI Baselines We compare HiOdex with Grad-CAM (Selvaraju et al., 2017), Grad-CAM++ (Chattopadhay et al., 2018), LayerCAM (Jiang et al., 2021), Eigen-CAM (Muhammad & Yeasin, 2020), G-CAME (Nguyen et al., 2024a), D-RISE (Petsiuk et al., 2021), D-CLOSE (Nguyen et al., 2024b), and the human-inspired FovEx (Panda et al., 2025). CAM-based methods use shared target layers, while perturbation-based methods retain their prescribed masking and segmentation procedures because these operations define the corresponding methods. HiOdex uses fixed detector-specific layer and fusion assignments, with all remaining parameters held constant across datasets and object instances. All methods explain the same retained detections under identical detector, input, and target settings, with results averaged over detections. For perturbation-based evaluation, predictions are reassociated with the same-class detection that best matches the original target. Complete detector configurations, baseline settings, and fairness controls are reported in App. B.

Metrics *Model-oriented explanation quality* (Sec. 5.1) is evaluated using six metrics. Deletion and Insertion (Petsiuk et al., 2018) measure how the selected prediction changes when highly attributed pixels are removed or restored, while OA (Zhang et al., 2021) summarizes the difference between the two scores. PG (Zhang et al., 2018) tests whether the attribution maximum lies inside the target

object, EBPG (Wang et al., 2020) measures the attribution mass assigned to that object, and Sparsity (Gomez et al., 2022) measures attribution concentration. Lower Deletion is better, while higher values are better for the other five measures. Generation cost is measured as the mean wall-clock time required to generate one attribution map on 1 NVIDIA A100 80 GB GPU.

Human-gaze alignment (Sec. 5.2) is evaluated using detections matched to gaze targets by class agreement and an intersection over union of at least 0.3. A same-class prediction is also accepted when its center lies inside the target box after each side has been expanded by 10%. Human fixations are pooled across observers after coordinate alignment. For COCO-Search18, only successful search trials are retained, and the forced initial central fixation is removed to avoid artificial center bias. The same protocol is applied to all explanation methods. Alignment is measured primarily using AUC-Judd (Judd et al., 2009), SIM (Bylinskii et al., 2018), and KL divergence (Bylinskii et al., 2018; Judd et al., 2012), which compare attribution with the whole-image fixation distribution. Higher AUC-Judd and SIM and lower KL divergence indicate better agreement. We additionally report Weighted Normalized Scanpath Saliency (WNSS) (Gide & Karam, 2016) and introduce OFCS and WDSIM as complementary object-level measures. Higher values indicate better agreement for all three measures. These metrics use only recorded human fixations. The model-generated fixations used internally by HiOdex never enter the gaze evaluation. App. B provides the complete metric definitions, gaze-processing procedure, and reference settings, while App. C.7 examines their parameter sensitivity.

5 Empirical Results

We compare HiOdex with the eight baselines using the fixed configuration reported in App. A.4. Controlled ablations of the HiOdex mechanisms themselves are reported in App. C.3.

5.1 Model-Oriented Explanation Quality

As shown in Table 1, HiOdex achieves the highest OA among the CAM-based methods and FovEx in all four detector and benchmark settings. Its OA remains within 0.0217 of the strongest perturbation-based result in every setting. Across all six metrics, HiOdex ranks among the top three methods in 22 of the 24 detector, benchmark, and metric combinations and never ranks below fourth. It also outperforms FovEx in all 24 comparisons. These results indicate a consistent balance among faithfulness, localization, and attribution concentration rather than strength on a single criterion. App. C.1 reports the ranking analysis, while App. D.1 shows that HiOdex also achieves the highest OA under both detectors on the domain-specific Substation Equipment dataset.

Qualitative Results Fig. 3 supplements with aligned cross-method comparisons under YOLOv12-N. For each selected detection, HiOdex distributes attribution across multiple evidence-bearing regions while remaining associated with the target object. The baselines range from highly concentrated responses to broader or spatially fragmented maps, illustrating why localization, concentration, and perturbation-based faithfulness can rank the same explanation differently. App. D.2 provides the corresponding Faster R-CNN comparison and more examples under both detectors.

Explanation Quality and Generation Cost Fig. 1 compares OA with the mean time required to generate one attribution map. HiOdex operates in the same subsecond regime as the CAM-based methods while remaining approximately two orders of magnitude faster than D-RISE and D-CLOSE and approximately one order of magnitude faster than FovEx. Within this runtime regime, it achieves the highest OA in all four detector and benchmark settings. It also remains within 0.0217 of the strongest perturbation-based result. HiOdex therefore provides a favorable quality-cost trade-off without requiring the repeated detector evaluations used by perturbation-based alternatives.

5.2 Human-Gaze Alignment

As our central contribution, we assess whether object-specific explanations produced by HiOdex align with human visual attention. As shown in Table 2, HiOdex achieves the highest AUC-Judd in all four detector and gaze-dataset settings, reaching 0.9265 and 0.7472 with YOLOv12-N and

Table 1: **Model-oriented explanation results.** Best values are shown in bold, and second-best values are underlined. † denotes human-inspired XAI methods.

		MS-COCO						PASCAL VOC					
		Faithfulness			Localization			Faithfulness			Localization		
Detector	Method	OA ↑	Del. ↓	Ins. ↑	PG ↑	EBPG ↑	Spars. ↑	OA ↑	Del. ↓	Ins. ↑	PG ↑	EBPG ↑	Spars. ↑
YOLOv12-N	HiOdex (Ours)†	.8106	.0530	.8636	.9910	.4956	58.7646	.7679	.0926	**.8605**	.9833	.6389	37.1371
	FovEx†	.7504	.0876	.8380	.3930	.2205	6.1270	.7235	.1096	.8331	.6230	.4060	6.1718
	G-CAME	.7940	.0610	.8550	**.9934**	.8670	446.8855	.7484	.0968	.8552	.9973	.9100	156.2385
	Grad-CAM	.6741	.0839	.7580	.5847	.4094	52.4731	.6018	.1164	.7182	.7077	.6162	51.0584
	Grad-CAM++	.6406	.1034	.7441	.4676	.1663	8.2482	.5325	.1751	.7076	.5823	.2914	9.7782
	Eigen-CAM	.5131	.1509	.6641	.2622	.1214	2.7122	.3222	.2574	.5796	.4375	.2266	2.1809
	LayerCAM	.3763	.1805	.5568	.9918	**.9378**	**486.4520**	.2392	.2591	.4983	**.9978**	**.9827**	**240.8627**
	D-RISE	**.8130**	.0570	**.8700**	.8604	.1786	5.8348	.7540	.0968	.8508	.9084	.3083	5.0117
	D-CLOSE	.8118	**.0203**	.8321	.9808	.2319	21.1636	**.7896**	**.0321**	.8217	.9904	.3411	14.2511
Faster R-CNN	HiOdex (Ours)†	.5712	.0278	.5990	**.9962**	.3561	33.3332	.5613	.0540	.6153	**.9977**	.5478	20.1983
	FovEx†	.4762	.0695	.5457	.2787	.1826	6.3428	.4779	.0898	.5677	.5841	.3757	5.7161
	G-CAME	.5202	.0279	.6081	.9857	**.6539**	**258.8158**	.5245	.0571	**.6216**	.9919	.7697	**65.6340**
	Grad-CAM	.3890	.0677	.4568	.3158	.1130	3.3798	.4353	.0815	.5168	.5339	.2484	3.4122
	Grad-CAM++	.2159	.1292	.3451	.1984	.1163	3.8201	.2452	.1531	.3983	.3897	.2815	5.5959
	Eigen-CAM	.0254	.2672	.2927	.0486	.0875	2.4223	.0439	.2688	.3127	.1375	.1975	2.0226
	LayerCAM	.5003	.0243	.5446	.8620	.6031	106.9391	.5056	.0485	.6041	.9348	**.7809**	46.3968
	D-RISE	.5751	.0368	**.6119**	.8115	.1802	6.9839	.5455	.0646	.6101	.9075	.2884	5.3340
	D-CLOSE	**.5773**	**.0177**	.5951	.9127	.2151	20.8601	**.5704**	**.0211**	.5915	.9833	.3531	16.5352

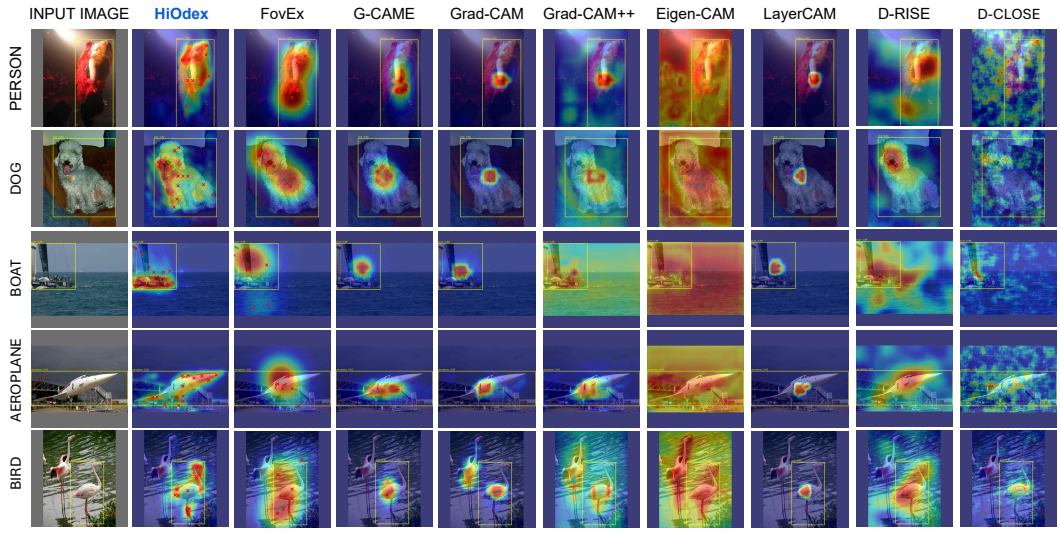

Figure 3: **Attribution comparison with YOLOv12-N.** Rows correspond to selected detections, and columns correspond to explanation methods. Yellow boxes indicate the selected detections. Red crosses indicate the model-generated fixations shown for HiOdex and FovEx.

0.9213 and 0.7491 with Faster R-CNN on COCO-Search18 and POET, respectively. The 0.0001 margin over FovEx with YOLOv12-N on POET constitutes an effective tie.

Across AUC-Judd, SIM, and KL divergence, HiOdex obtains the best average rank among the nine evaluated methods. It retains this position when OFCS, WNSS, and WDSIM are included and ranks among the top three methods in 17 of the 24 gaze comparisons. The broader metric set is less uniform on POET, where different measures favor different spatial distributions, which is further examined in App. C.1.

Table 2: **Human-gaze alignment results.** Best values are shown in bold, and second-best values are underlined. † denotes human-inspired XAI methods.

		COCO-Search18						POET					
		Image-level			Object-level			Image-level			Object-level		
Detector	Method	AUC-Judd ↑	SIM ↑	KL ↓	OFCS ↑	WNSS ↑	WDSIM ↑	AUC-Judd ↑	SIM ↑	KL ↓	OFCS ↑	WNSS ↑	WDSIM ↑
YOLOv12-N	HiOdex (Ours)†	**.9265**	.3755	2.2911	.8226	6.1669	.7914	**.7472**	.4500	1.9951	.4770	1.2403	.7134
	FovEx†	.8523	.1424	2.5816	.7876	1.7483	.7841	.7471	**.5125**	**.9237**	.5802	1.2784	.7425
	G-CAME	.8966	**.4209**	5.5466	.5670	**7.3346**	.7071	.7410	.3734	6.0873	.3757	1.3554	.6280
	Grad-CAM	.9213	.4000	3.6508	.7231	6.9199	.7647	.7337	.3646	9.0222	.3280	**1.4778**	.6707
	Grad-CAM++	.8912	.1410	2.4780	.7585	4.1450	.7871	.6909	.4192	1.1745	.5094	.9995	.7581
	Eigen-CAM	.8471	.0856	3.1070	.8015	2.4260	.7909	.6203	.3860	1.1251	**.6230**	.6078	.7506
	LayerCAM	.7634	.3321	13.5329	.5285	6.7797	.6497	.6098	.1455	20.7298	.2202	1.0319	.5103
	D-RISE	.9043	.1408	2.4600	**.8862**	3.5014	**.8081**	.7106	.4444	.9542	.5901	1.2014	**.7642**
	D-CLOSE	.9004	.2397	**1.8703**	.7336	6.5758	.7771	.6924	.4458	1.0027	.3933	1.4196	.7640
Faster R-CNN	HiOdex (Ours)†	**.9213**	.3275	2.3415	.8277	5.4106	.7985	**.7491**	.4796	1.7671	**.5320**	**1.4669**	.7137
	FovEx†	.8333	.1411	2.6485	.7042	1.4766	.7793	.7274	**.4940**	**.9665**	.4931	1.1349	.7345
	G-CAME	.9050	**.4144**	4.4086	.7282	6.7156	.7763	.7372	.4193	4.8010	.4919	1.3945	.6635
	Grad-CAM	.7385	.0864	3.1211	.7154	1.4174	.7884	.6832	.4188	1.0657	.4945	1.0346	**.7669**
	Grad-CAM++	.4910	.0758	3.4448	.5578	-.1782	.7735	.5865	.4021	2.2389	.3882	.5061	.7499
	Eigen-CAM	.1810	.0572	3.7921	.5134	-1.6199	.7689	.3962	.3595	1.2776	.4877	-.6788	.7432
	LayerCAM	.8732	.3853	6.6815	.8277	5.3409	.7653	.6408	.3619	10.8254	.5030	1.1954	.6186
	D-RISE	.9003	.1548	2.3894	**.8533**	3.6803	**.8009**	.6899	.4361	.9992	.5058	1.0515	.7520
	D-CLOSE	.8853	.2412	**1.8861**	.7075	**6.7484**	.7859	.6819	.4416	1.0253	.3542	1.3688	.7627

HiOdex uses neither recorded fixations nor gaze supervision. Its agreement with human observations therefore emerges from detector-derived evidence and the spatial organization imposed by its fixation process. The results show that HiOdex combines competitive detector faithfulness, efficient explanation generation, and consistent human-gaze alignment.

6 Conclusion

We introduced HiOdex, a gaze-free, object-specific explainer that embeds human-inspired fixation behavior within detector-feature attribution. It derives model-generated fixations from activation-gradient evidence and uses their spatial envelope to guide channel weighting, while local evidence, object-context gating, and multi-layer fusion preserve spatial detail. HiOdex requires only one forward and one backward pass, and recorded human gaze remains independent evaluation evidence. Across YOLOv12-N and Faster R-CNN on MS-COCO and PASCAL VOC, HiOdex achieves the highest OA among the CAM-based methods and FovEx in all four settings, ranks among the top three methods in 22 of the 24 model-oriented comparisons, and operates in the subsecond regime. It also obtains the highest AUC-Judd score in all four gaze settings, with one effective tie. Controlled ablations show that object-context gating and high-resolution evidence materially support localization, while the fixation process reaches diminishing returns after the first few selections. Together, these findings show that human-inspired spatial organization can support efficient object-detection explanations without gaze supervision or repeated detector evaluations. The observed gaze agreement should nonetheless be interpreted as behavioral correspondence rather than evidence of human utility, which remains to be established through sequence-aware and task-based evaluation.

AI use statement

In this work, we used generative AI tools solely to polish the manuscript's writing. All AI-assisted text was reviewed and revised by the authors. We take responsibility for the final content of this work, including any text produced with the aid of generative AI.

Ethics statement

Explaining model predictions is important for the transparent and responsible use of AI systems. We evaluate HiOdex on established public image and human-gaze datasets. The recorded gaze data serves only as independent evaluation evidence for assessing agreement between model explanations and observed human visual attention, rather than as training supervision.

Reproducibility statement

To support reproducibility, we document the method and experimental protocol throughout the paper and appendices. Sec. 3 defines HiOdex, and App. A provides the full algorithm, notation, and fixed configuration. Sec. 4 describes the experimental design, while App. B specifies the data-processing pipeline, detector and baseline configurations, gaze-matching procedure, metric implementations, and computational protocol. Sec. 5 reports the main comparisons, while App. C documents the ranking analysis, controlled ablations, full evaluation grids, and parameter-sensitivity analyses. App. D provides the protocol and results for the additional domain-specific evaluation.

References

Philine Lou Bommer, Marlene Kretschmer, Anna Hedström, Dilyara Bareeva, and Marina M-C Höhne. Finding the right xai method—a guide for the evaluation and ranking of explainable ai methods in climate science. *Artificial Intelligence for the Earth Systems*, 3(3):e230074, 2024.

Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847. IEEE, 2018.

Yupei Chen, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. Coco-search18 fixation dataset for predicting goal-directed attention control. *Scientific reports*, 11(1):8776, 2021.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pp. 226–231, 1996.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Milind S Gide and Lina J Karam. A locally weighted fixation density-based metric for assessing the quality of visual saliency predictions. *IEEE Transactions on Image Processing*, 25(8):3852–3861, 2016.

Andreas Anael Pereira Gomes, Francisco Itamarati Secolo Ganacim, Fabiano Gustavo Silveira Magrin, Nara Bobko, Leonardo Göbel Fernandes, Anselmo Pombeiro, and Eduardo Félix Ribeiro Romaneli. A semantically annotated 15-class ground truth dataset for substation equipment to train semantic segmentation models. *Data*, 8(7):118, 2023.

Tristan Gomez, Thomas Fréour, and Harold Mouchère. Metrics for saliency map evaluation of deep learning explanation methods. In *International conference on pattern recognition and artificial intelligence*, pp. 84–95. Springer, 2022.

Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE transactions on image processing*, 30:5875–5888, 2021.

Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pp. 2106–2113. IEEE, 2009.

Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. Technical Report MIT-CSAIL-TR-2012-001, Massachusetts Institute of Technology, January 2012.

Rémi Kazmierczak, Steve Azzolin, Eloïse Berthier, Anna Hedström, Patricia Delhomme, David Filliat, Nicolas Bousquet, Goran Frehse, Massimiliano Mancini, Baptiste Caramiaux, et al. Benchmarking xai explanations with human-aligned evaluations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 37491–37500, 2026.

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.

Guoyang Liu, Jindi Zhang, Antoni B Chan, and Janet H Hsiao. Human attention guided explainable artificial intelligence for computer vision models. *Neural Networks*, 177:106392, 2024.

Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 international joint conference on neural networks (IJCNN)*, pp. 1–7. IEEE, 2020.

Loc Phuc Truong Nguyen, Hung Truong Thanh Nguyen, and Hung Cao. Odexai: A comprehensive object detection explainable ai evaluation. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)*, pp. 118–133. Springer, 2025.

Quoc Khanh Nguyen, Truong Thanh Hung Nguyen, Vo Thanh Khang Nguyen, Van Binh Truong, Tuong Phan, and Hung Cao. Efficient and concise explanations for object detection with gaussian-class activation mapping explainer. *arXiv preprint arXiv:2404.13417*, 2024a.

Truong Thanh Hung Nguyen, Vo Thanh Khang Nguyen, Quoc Khanh Nguyen, Quoc Hung Cao, et al. Towards better explanations for object detection. In *Asian Conference on Machine Learning*, pp. 1385–1400. PMLR, 2024b.

Mahadev Prasad Panda, Matteo Tiezzi, Martina Vilas, Gemma Roig, Bjoern M Eskofier, and Dario Zanca. Fovex: Human-inspired explanations for vision transformers and convolutional neural networks. *International journal of computer vision*, 133(10):7437–7459, 2025.

Dim P Papadopoulos, Alasdair DF Clarke, Frank Keller, and Vittorio Ferrari. Training object class detectors from eye tracking data. In *European conference on computer vision*, pp. 361–376. Springer, 2014.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11438–11447. IEEE, 2021.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 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 *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2591–2600, 2019.

Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *Advances in neural information processing systems*, 38:78433–78457, 2026.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *2020 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 111–119. IEEE, 2020.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126 (10):1084–1102, 2018.

Qinglong Zhang, Lu Rao, and Yubin Yang. Group-cam: Group score-weighted visual explanations for deep convolutional networks. *arXiv preprint arXiv:2103.13859*, 2021.

Chenyang Zhao, Janet H Hsiao, and Antoni B Chan. Gradient-based instance-specific visual explanations for object specification and object discrimination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9):5967–5985, 2024.

A TECHNICAL BACKGROUND

HiOdex combines evidence-driven fixation selection, fixation-envelope-guided gradient pooling, position-wise local evidence, object-context refinement, and multi-layer fusion. This appendix shows the method's technical distinctions, computational dependencies, inference procedure, and fixed evaluation configuration.

A.1 DESIGN DISTINCTIONS

HiOdex is a gaze-free, object-specific explanation method that directly incorporates model-generated fixation behavior into detector-feature attribution. Its closest human-inspired reference is FovEx (Panda et al., 2025). Both methods derive fixations from model evidence rather than recorded human gaze. FovEx uses these fixations to transform the input iteratively and aggregates explanations across the resulting foveated views. HiOdex instead constructs a fixation envelope within the selected detection and uses this envelope to determine where detection-score gradients are pooled into class activation mapping channel weights. The fixation process therefore shapes the attribution during its construction without modifying the input image or requiring additional detector evaluations.

HiOdex also differs from G-CAME (Nguyen et al., 2024a), which applies a target-centered spatial prior to the resulting saliency map. In HiOdex, detector evidence determines the spatial support for gradient pooling before the Foveal CAM is formed. A parallel Local Evidence Map retains position-wise activation-gradient structure that channel pooling may suppress. The two branches are subsequently integrated, refined around the selected object, and fused across feature levels.

Recorded human gaze is not used at any stage of explanation generation. The model-generated fixations $f_m^{(l)}$ used by HiOdex are therefore distinct from the recorded human fixations used exclusively for evaluation. This separation preserves human gaze as an independent behavioral reference rather than an optimization signal.

A.2 COMPUTATIONAL DEPENDENCIES

HiOdex explains a selected detection $\tau = (b_\tau, c_\tau, s_\tau)$, where s_τ is the differentiable score associated with the matched prediction before non-maximum suppression. One forward pass and one backward pass provide the activations $A^{(l)}$ and detection-score gradients $G^{(l)}$ for every target layer $l \in \mathcal{L}$. All subsequent operations act on these cached feature maps. The complete dependency structure is

$$
\begin{aligned}
&\left(A^{(l)}, G^{(l)}\right) \longrightarrow \left(E^{(l)}, \widetilde{E}^{(l)}\right) \longrightarrow \left\{f_m^{(l)}, Q_m^{(l)}, m_m^{(l)}, \pi_m^{(l)}\right\}_{m=1}^{K} \longrightarrow Q^{(l)} \longrightarrow H_F^{(l)}, \\
&\left(A^{(l)}, G^{(l)}\right) \longrightarrow \widehat{G}^{(l)} \longrightarrow H_L^{(l)}, \\
&\left(H_F^{(l)}, H_L^{(l)}\right) \longrightarrow M^{(l)} \longrightarrow \bar{M}^{(l)} \longrightarrow H.
\end{aligned}
$$

Normalization to $[0, 1]$ is applied at the stages specified in Sec. 3. Gaussian smoothing is measured in feature cells for layer-level maps and in pixels for image-space maps. Bilinear interpolation maps each layer attribution to the input resolution. The numerical constant ε stabilizes the denominators used for gradient normalization, fixation weighting, and envelope-guided pooling. Table A1 summarizes the computational dependencies and links each stage to its defining section or equation.

A.3 COMPLETE HIODEX PROCEDURE

Algorithm A1 presents the complete inference procedure. All quantities follow the definitions in Sec. 3 and Table A1.

All target layers reuse the activations and gradients obtained from the same forward pass and backward pass. HiOdex therefore requires no further detector evaluation after backpropagation. The remaining computation consists of feature-map processing, fixation selection, smoothing, resizing, object-context gating, and layer fusion.

Table A1: Computational dependencies in HiOdex.

Stage	Quantities	Operation, role, and pointer
Detector evidence	$A^{(l)}, G^{(l)}$	Activations and gradients associated with the selected detection score (Sec. 3.1).
Positive evidence	$E^{(l)}, \widetilde{E}^{(l)}$	Combines positive activations and gradients, then suppresses isolated feature-cell responses (Eq. (1)).
Fixation selection	$f_m^{(l)}, Q_m^{(l)}, W_m^{(l)}$	Selects successive evidence maxima and applies object-scaled inhibition of return (Sec. 3.2).
Fixation envelope	$m_m^{(l)}, \pi_m^{(l)}, Q^{(l)}$	Weights fixation windows by captured evidence and aggregates them within the selected detection (Eq. (4)).
Foveal CAM	$\alpha_k^{(l)}, H_F^{(l)}$	Pools gradients within the fixation envelope and forms the channel-weighted attribution (Eq. (5)).
Local evidence	$\widehat{G}_k^{(l)}, H_L^{(l)}$	Preserves position-wise positive activation-gradient structure (Eq. (6)).
Layer attribution	$M^{(l)}, R, \bar{M}^{(l)}$	Integrates the two branches, resizes the resulting map, and suppresses responses distant from the selected object (Eq. (7)).
Final explanation	H	Fuses the refined layer maps according to feature resolution and normalizes the result (Eq. (7)).

Algorithm A1 Complete HiOdex inference procedure.

Require: Detector D_θ, image x, selected detection $\tau = (b_\tau, c_\tau, s_\tau)$, target layers $\mathcal{L}$, scalar parameters $\delta, K, \gamma, \beta, \rho$, and fusion weights $\boldsymbol{\omega}$
Ensure: HiOdex attribution map H

```
Run one forward pass and one backward pass to obtain {A^(l), G^(l)}_{l∈L}
for each l ∈ L do
    Compute E^(l), smooth it into Ẽ^(l), and project b_τ onto layer l
    Initialize W_1^(l) using the evidence inside b_τ^(l)
    for m = 1 to K do
        Select f_m^(l) ← arg max W_m^(l) and construct Q_m^(l)
        Compute m_m^(l) and update W_{m+1}^(l) ← W_m^(l) ⊙ (1 − Q_m^(l))
    end for
    Compute {π_m^(l)}_{m=1}^K and the fixation envelope Q^(l)
    Compute {α_k^(l)}_{k=1}^{C_l} and the Foveal CAM H_F^(l)
    Compute {Ĝ_k^(l)}_{k=1}^{C_l} and the Local Evidence Map H_L^(l)
    Integrate the branches into M^(l), resize the map, and apply R to obtain M̄^(l)
end for
H ← Norm(Σ_{l∈L} ω_l M̄^(l))
return H
```

A.4 Fixed Configuration

Table A2 reports the reference configuration used throughout the evaluation. The scalar parameters are shared across detectors, datasets, images, and selected objects. The fusion policy is fixed before evaluation while accounting for the feature hierarchy of each detector.

The detector-specific target layers and fusion assignments are reported in Table B2. The reference configuration is treated as a stable operating point rather than a uniquely optimal setting. App. C.3 summarizes the mechanism ablations, while Appendices C.4–C.7 report the complete parameter sweeps and qualitative sensitivity analyses.

Table A2: Fixed HiOdex configuration.

Symbol	Value or rule	Description
δ	0.40	Fixation-window width relative to the corresponding side of the projected bounding box
K	3	Number of model-generated fixations selected at each target layer
γ	0.04	Attribution-smoothing width relative to $\min(H_l, W_l)$
β	0.50	Mixture weight between the Foveal CAM and Local Evidence Map
ρ	0.60	Object-context gate width relative to the corresponding side of the selected bounding box
$\boldsymbol{\omega}$, YOLOv12-N	$(0.55, 0.35, 0.10)$	Fusion weights for the P3, P4, and P5 detection-input layers, ordered from highest to lowest spatial resolution
$\boldsymbol{\omega}$, Faster R-CNN R50-FPN	Uniform	Equal fusion across the selected feature-pyramid lateral and output blocks

B Experimental Protocol, Metrics, and Reproducibility

This appendix provides the complete evaluation protocol supporting Sections 4 and 5. It specifies the evaluation data, detector configurations, explanation targets, gaze processing, baseline settings, metric definitions, and computational protocol. All methods are evaluated on matched detections under the same preprocessing, target definition, and aggregation procedure.

B.1 Evaluation Data and Detector Configurations

Table B1 summarizes the role, scale, category coverage, and observer protocol of each evaluation dataset.

Table B1: Datasets and evaluation scope.

Dataset	Evaluation role	Images	Classes	Observers	Observer task
MS-COCO	Detection benchmark	1,000	80	–	–
COCO-Search18	Gaze reference paired with MS-COCO	6,202	18	10	Search for a cued target category
PASCAL VOC	Detection benchmark	1,000	20	–	–
POET	Gaze reference paired with PASCAL VOC	6,270	10	5	Identify which object class is present

MS-COCO (Lin et al., 2014) provides instance-level annotations for 80 object categories in complex natural scenes. COCO-Search18 (Chen et al., 2021) contains 6,202 MS-COCO images and approximately 300,000 fixations from ten observers searching for targets from 18 categories. PASCAL VOC (Everingham et al., 2010) provides bounding-box annotations for 20 categories. POET (Papadopoulos et al., 2014) augments 6,270 images from ten PASCAL VOC 2012 categories with eye movements from five observers. COCO-Search18 presents a target category before each search trial, whereas POET asks observers to identify which of ten object classes is present. POET records one second of viewing and therefore provides sparser fixation evidence than COCO-Search18.

We evaluate the one-stage YOLOv12-N (Tian et al., 2026) and the two-stage Faster R-CNN (Ren et al., 2016) with a ResNet-50 feature-pyramid backbone. Both detectors receive 640×640 letterboxed inputs produced by the same preprocessing routine. This establishes a common coordinate frame for detector outputs, attribution maps, and model-generated fixations. Predictions are retained at a confidence threshold of 0.25, followed by non-maximum suppression with an intersection-over-union threshold of 0.45. For Faster R-CNN, these criteria are applied after the internal score filtering and class-wise suppression implemented by `torchvision`.

Detector outputs are mapped to a common category vocabulary. Contiguous PASCAL VOC indices are converted using a fixed offset, while MS-COCO categories are mapped by name to the COCO-80 ordering. Predictions from unmapped categories are excluded. The detectors use COCO-pretrained checkpoints, with the PASCAL VOC models fine-tuned from their corresponding pretrained weights. HiOdex requires only that the selected detection score remain differentiable with respect to the selected spatial feature layers. Table B2 reports the detector-specific attribution layers and the corresponding HiOdex fusion policy.

Table B2: Detector-specific attribution configurations.

Detector	**Paradigm**	**Attribution target layers**	**HiOdex fusion**
YOLOv12-N	One-stage	P3, P4, and P5 detection-input layers at `model.14`, `model.17`, and `model.20`	$(0.55, 0.35, 0.10)$
Faster R-CNN R50-FPN	Two-stage	Feature-pyramid lateral and output convolutional blocks	Uniform

For YOLOv12-N, attribution is extracted from the P3, P4, and P5 detection-input layers. Their explanation maps are fused with weights $(0.55, 0.35, 0.10)$. For Faster R-CNN, attribution is extracted from the lateral and output convolutional blocks of the feature pyramid and fused uniformly. Every CAM-based comparator uses the same detector-specific target layers. Consequently, differences between these methods arise from their attribution procedures rather than from access to different representations.

B.2 Explanation Targets and Paired Evaluation

Each retained detection is explained independently. An image with several retained detections therefore contributes several explanations, and model-oriented results are averaged over detections rather than images. The explanation target is the differentiable pre-suppression score matched to the selected detection. All gradient-based methods explain this same scalar target.

Model-oriented evaluation uses one fixed 1,000-image subset from each detection benchmark. The same images, detector checkpoints, retained detections, and explanation targets are used for every method, yielding paired comparisons. Human-gaze evaluation uses the complete COCO-Search18 and POET datasets after applying the matching criteria in App. B.3.

The subset restriction reflects the computational cost of perturbation-based evaluation and baselines. At 640×640 resolution, a Deletion or Insertion curve with 1,000-pixel perturbation steps requires approximately 411 detector evaluations per explanation. D-RISE (Petsiuk et al., 2021) uses 2,000 masked evaluations. D-CLOSE (Nguyen et al., 2024b) likewise uses 2,000 masks distributed across five segmentation levels.

B.3 Gaze Matching and Coordinate Processing

A predicted detection is associated with a gaze target when the predicted and annotated classes agree and their boxes have an intersection over union of at least 0.3. When this criterion is not satisfied, a same-class prediction is accepted if its center lies inside the target box after each side has been expanded by 10%, subject to a minimum expansion of two pixels. This secondary rule retains valid detections whose predicted extent differs from the annotation. Predictions satisfying neither condition are excluded from gaze evaluation. The same matching procedure is applied to every explanation method.

Category names are canonicalized across datasets, including airplane and aeroplane, motorcycle and motorbike, and couch and sofa. Object-level gaze measures use the predicted box of the explained detection rather than its ground-truth box. Human fixations are pooled without observer-specific weighting.

For COCO-Search18, only successful search trials are retained. The forced initial central fixation is removed to prevent an artificial center bias. A retained trial is not required to contain a fixation inside the target box. The recorded 1680×1050 display coordinates are mapped to image coordinates using aspect-preserving scaling and the corresponding centering offsets. Fixations outside the displayed

image are discarded. Detection-to-target matching is performed in the letterboxed coordinate frame, while gaze metrics are computed in the original-image coordinate frame.

For the cluster-based metrics, pooled fixations associated with each matched object are clustered using Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996). The reference configuration uses a neighborhood radius $\mathrm{Eps} = 1^\circ$ of visual angle and $\mathrm{MinPts} = 3$. Fixations classified as noise are excluded. The same fixation clusters are used for all compared methods, so cluster eligibility and exclusions do not depend on the attribution map.

B.4 Baselines and Fairness Controls

We compare HiOdex with Grad-CAM (Selvaraju et al., 2017), Grad-CAM++ (Chattopadhay et al., 2018), LayerCAM (Jiang et al., 2021), Eigen-CAM (Muhammad & Yeasin, 2020), G-CAME (Nguyen et al., 2024a), D-RISE (Petsiuk et al., 2021), D-CLOSE (Nguyen et al., 2024b), and the human-inspired FovEx (Panda et al., 2025). Every method receives the same detector input and explains the same retained detection. CAM-based methods access the same detector-specific feature layers listed in Table B2. Perturbation-based methods retain their prescribed masking and segmentation procedures because these operations define the corresponding methods. Table B3 reports the method-specific settings used in all comparisons.

Table B3: Baseline configurations.

Method	**Family**	**Evaluation configuration**
Grad-CAM, Grad-CAM++, LayerCAM, Eigen-CAM	CAM-based	Shared detector-specific layers from Table B2, with the published method-specific aggregation procedure.
G-CAME	CAM-based	Shared target layers, uniform layer aggregation, and a confidence-aware anisotropic Gaussian with a spread divisor from 3.0 to 6.0 and a minimum spread of 0.5 pixels.
D-RISE	Perturbation-based	2,000 masks, mask probability 0.5, and 8×8 mask resolution.
D-CLOSE	Perturbation-based	SLIC segmentation at 150, 300, 600, 1,200, and 2,400 segments, with 400 masks per level, mask probability 0.5, and resize offset 2.2.
FovEx	Human-inspired	`RR=True`, $\sigma_f = 0.15$, $\beta = 0.1$, $N = 10$, $\mathrm{OS} = 20$, $\sigma_b = 10$, $\mathrm{BFS} = 41$, and $\lambda = 0.1$.
HiOdex (Ours)	Human-inspired	Fixed configuration from Table A2, without detector-specific, dataset-specific, or instance-specific tuning.

*FovEx hyperparameters follow the notation of its original work. Its β, σ_f, σ_b, N and λ are unrelated to the HiOdex parameters in Table A2.

HiOdex uses one fixed configuration across detectors, datasets, and object instances. Its complete method configuration is reported in Table A2. No result is selected through detector-specific or dataset-specific tuning.

B.5 Model-Oriented Metrics

Deletion and Insertion (Petsiuk et al., 2018) measure whether the most salient pixels causally affect the selected detection. Deletion progressively replaces pixels with zeros in descending attribution order. Insertion begins from an image blurred with a 51×51 Gaussian kernel and restores pixels in the same order. Both procedures perturb 1,000 pixels at each step and summarize the resulting target-score curve by its trapezoidal area. Lower Deletion and higher Insertion indicate better faithfulness.

After each perturbation step, the evaluation selects the same-class prediction with the greatest intersection over union relative to the original target box. When class-score vectors are available, its response combines this overlap with a Gaussian-kernel similarity between the original and perturbed class-score vectors. The kernel is applied to cosine distance with width 0.25. When class-score vectors are unavailable, the response is the product of the overlap and the detection confidence. This matching rule prevents a method from receiving credit when the retained prediction shifts to another object or class.

Let $S : \Omega \to [0, 1]$ denote the normalized attribution map, let $b_\tau \subseteq \Omega$ denote the target box, and let $p^* = \arg\max_{p\in\Omega} S(p)$. Over-All (OA) (Zhang et al., 2021), Pointing Game (PG) (Zhang et al., 2018), Energy-Based Pointing Game (EBPG) (Wang et al., 2020), and Sparsity (Gomez et al., 2022) are computed as

$$\begin{aligned} \mathrm{OA} &= \mathrm{Insertion} - \mathrm{Deletion}, \\ \mathrm{PG} &= \mathbf{1}[p^* \in b_\tau], \\ \mathrm{EBPG} &= \frac{\sum_{p\in b_\tau} S(p)}{\sum_{p\in\Omega} S(p)}, \\ \mathrm{Sparsity} &= \frac{\max_{p\in\Omega} S(p)}{|\Omega|^{-1} \sum_{p\in\Omega} S(p)}. \end{aligned} \tag{B1}$$

Higher values are better for all four measures. OA combines the two perturbation directions. PG tests whether the attribution maximum lies inside the target object. EBPG measures the fraction of attribution mass assigned to the object. Sparsity measures attribution concentration and is interpreted jointly with faithfulness and localization because a concentrated map is not necessarily correct.

B.6 Human-Gaze Metrics

The symbols introduced in this subsection describe the gaze metrics and are independent of the method notation in Section 3.

AUC-Judd (Judd et al., 2009), Similarity (SIM) (Bylinskii et al., 2018), and Kullback-Leibler (KL) divergence (Judd et al., 2012) compare an attribution map with the whole-image human fixation distribution. AUC-Judd treats recorded fixations as positive locations and evaluates how well attribution values separate them from non-fixated locations. SIM measures the histogram intersection between normalized attribution and fixation-density maps. KL divergence measures their distributional discrepancy. AUC-Judd and SIM are higher when better, whereas KL divergence is lower when better. Fixation-density maps use a Gaussian width of $\sigma = 1^\circ$ of visual angle.

Whole-image measures do not identify the particular prediction being explained when an image contains several objects. To complement these image-level metrics, we use the established Weighted Normalized Scanpath Saliency (WNSS) (Gide & Karam, 2016) and propose Object-Focused Cluster Score (OFCS) and Weighted Distribution Similarity (WDSIM) for object-level comparisons based on human-consensus fixation clusters. These measures use only recorded human fixations. The model-generated fixations used internally by HiOdex never enter their computation. Table B4 defines the notation shared by the three object-level gaze metrics.

Let $C_1, \ldots, C_{N_C}$ denote the valid DBSCAN clusters associated with an explained detection, with $n_k = |C_k|$. Cluster k receives the normalized consensus weight

$$\eta_k = \frac{n_k^2}{\sum_{j=1}^{N_C} n_j^2}. \tag{B2}$$

The squared cluster size follows directly from applying the published cluster-size weight to every fixation and then grouping the fixation-level sum by cluster. It is not interpreted as literal pairwise observer agreement because one observer may contribute several fixations to the same cluster.

WNSS For WNSS, the attribution map is standardized over the valid image domain:

$$z(p) = \frac{S(p) - \mu_S}{\sigma_S}, \qquad \bar{z}_k = \frac{1}{n_k} \sum_{g_i \in C_k} z(g_i), \tag{B3}$$

where μ_S and σ_S are the mean and standard deviation of S over Ω. WNSS is then

$$\mathrm{WNSS} = \sum_{k=1}^{N_C} \eta_k \bar{z}_k. \tag{B4}$$

Higher values indicate that attribution is stronger at dense human-consensus locations.

Table B4: Notation for object-level gaze metrics.

Symbol	Meaning
Ω	Valid image domain
p	Image pixel
$S(p)$	Normalized attribution at pixel p
g_i	Recorded human fixation event
$f_m^{(l)}$	Model-generated fixation m at feature level l
b_τ	Predicted box of the explained detection
C_k	Valid DBSCAN fixation cluster k
n_k	Number of fixations in C_k
N_C	Number of valid fixation clusters
η_k	Consensus weight of cluster k
Eps	DBSCAN neighborhood radius
MinPts	Minimum DBSCAN cluster size
$\mathcal{R}_k$	Disjoint image region associated with cluster k
P_k	Human-consensus distribution over clusters
Q_k	Attribution distribution over clusters
$\bar{S}_k$	Mean attribution over region $\mathcal{R}_k$
Φ	Consensus-weighted alignment response (OFCS)
Ψ	Normalized cluster-participation ratio (OFCS)

OFCS OFCS evaluates both the attribution response at human-consensus locations inside the predicted box and the extent to which that response covers the valid clusters. A fixation contributes its attribution value when it lies inside b_τ and contributes zero otherwise:

$$r_i = \mathbf{1}[g_i \in b_\tau] S(g_i), \qquad \bar{r}_k = \frac{1}{n_k} \sum_{g_i \in C_k} r_i. \tag{B5}$$

The consensus-weighted alignment response is

$$\Phi = \sum_{k=1}^{N_C} \eta_k \bar{r}_k. \tag{B6}$$

Cluster coverage is measured by the normalized participation ratio

$$\Psi = \begin{cases} \dfrac{\left(\sum_{k=1}^{N_C} \bar{r}_k\right)^2}{N_C \sum_{k=1}^{N_C} \bar{r}_k^2}, & \displaystyle\sum_{k=1}^{N_C} \bar{r}_k^2 > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{B7}$$

OFCS combines alignment and coverage through their geometric mean:

$$\mathrm{OFCS} = \sqrt{\Phi \Psi}. \tag{B8}$$

OFCS lies in $[0, 1]$, is defined for $N_C \geq 1$, and is higher when better. A cluster outside the predicted box remains in the coverage calculation with zero response. Its presence therefore lowers the score consistently for every method evaluated on that detection.

WDSIM WDSIM measures whether human fixations and attribution distribute their mass similarly across consensus clusters. Its human-side distribution is

$$P_k = \eta_k = \frac{n_k^2}{\sum_{j=1}^{N_C} n_j^2}. \tag{B9}$$

Each cluster defines an image region

$$\mathcal{R}_k = \left\{ p \in \Omega \;\middle|\; \min_{g_i \in C_k} \|p - g_i\|_2 \leq \mathrm{Eps} \right\}. \tag{B10}$$

When regions overlap, each shared pixel is assigned to its nearest fixation cluster, producing disjoint regions. Reusing Eps avoids introducing another spatial parameter. The regional attribution response and its normalized distribution are

$$\bar{S}_k = \frac{1}{|\mathcal{R}_k|} \sum_{p \in \mathcal{R}_k} S(p), \qquad Q_k = \frac{\bar{S}_k}{\sum_{j=1}^{N_C} \bar{S}_j}. \tag{B11}$$

WDSIM applies histogram intersection to the human and attribution distributions:

$$\mathrm{WDSIM} = \begin{cases} \sum_{k=1}^{N_C} \min(P_k, Q_k), & \sum_{j=1}^{N_C} \bar{S}_j > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{B12}$$

WDSIM lies in $[0, 1]$, is defined for $N_C \geq 2$, and is higher when better. Regional means prevent large regions from receiving greater weight solely because they contain more pixels. OFCS and WDSIM therefore answer complementary questions. OFCS tests whether human-consensus locations receive attribution within the selected object, while WDSIM tests whether attribution is distributed across those locations in proportions consistent with human consensus.

Objects without the minimum number of valid clusters are excluded only from the metric for which the requirement is not met. OFCS and WNSS require at least one cluster, while WDSIM requires at least two. Because cluster construction depends only on recorded fixations, these exclusions are identical across explanation methods and preserve paired comparisons. The complete parameter-sensitivity analysis is reported in App. C.7.

B.7 Computational Protocol

Computational cost is measured as the mean wall-clock time required to generate one attribution map on an otherwise idle NVIDIA A100 80 GB GPU. Every method uses a 640×640 input. Image loading, model construction, disk access, and metric evaluation are excluded, so the measurement isolates explanation generation.

Runtime is averaged over explained detections within each detector and dataset setting. The evaluation uses fixed detector checkpoints, fixed evaluation subsets, and matched detections. The reported comparisons therefore characterize explanation behavior for the evaluated models rather than variability across detector-training runs. When a gaze metric is undefined because an object does not satisfy its fixation-cluster requirement, the object is excluded from that metric rather than assigned an artificial numerical value.

C Ranking Analysis, Mechanism Ablations, and Parameter Sensitivity

This appendix provides the extended analysis supporting the principal results in Sec. 5. It first analyzes the method rankings behind Tables 1 and 2 and the associated computational cost. It then specifies the ablation protocol, summarizes the four principal mechanisms, and reports the complete parameter sweeps for fixation behavior, attribution construction, object-context gating, and multi-layer fusion. It closes with the sensitivity of the gaze metrics to their own clustering and density-estimation settings. All comparisons follow the paired evaluation protocol in App. B.

C.1 Ranking Analysis and Computational Cost

Model-oriented rankings. In Table 1, HiOdex ranks among the top three methods for both Deletion and Insertion in all four detector and benchmark settings. It achieves the highest Insertion score with YOLOv12-N on PASCAL VOC and the highest PG score with Faster R-CNN on both benchmarks. Under YOLOv12-N, its PG rank is third on MS-COCO and fourth on PASCAL VOC, where the leading values exceed its scores by only 0.24 and 1.45 percentage points. HiOdex also ranks third on EBPG in all four settings and third on Sparsity in three settings, falling to fourth only with YOLOv12-N on PASCAL VOC.

The criteria do not always agree. LayerCAM achieves the highest EBPG and Sparsity with YOLOv12-N on both benchmarks, yet produces the lowest OA in the same settings. Concentrated attribution is therefore not sufficient for perturbation-based faithfulness. Against the 22 top-three placements of HiOdex reported in Sec. 5.1, G-CAME reaches the top three 17 times, while D-CLOSE and D-RISE do so 11 and 10 times and fall as low as sixth and eighth, respectively. HiOdex also exceeds FovEx in every model-oriented comparison. Its largest margin is 0.7175 on PG with Faster R-CNN and MS-COCO.

Human-gaze rankings. In Table 2, AUC-Judd, SIM, and KL divergence provide the primary whole-image comparison, while OFCS, WNSS, and WDSIM provide complementary object-level evidence. On COCO-Search18, HiOdex exceeds FovEx on all six metrics under both detectors and appears among the top three in 11 of the 12 comparisons. Results on POET depend more strongly on the detector and metric. With Faster R-CNN, HiOdex achieves the highest OFCS, WNSS, and AUC-Judd scores and the second-highest SIM. With YOLOv12-N, it retains the highest AUC-Judd and second-highest SIM but ranks lower on the remaining measures.

Metric rankings and dataset effects. Fig. C1 summarizes the method ranks across the six human-gaze metrics and reveals a more consistent pattern on COCO-Search18 than on POET. This difference is compatible with their distinct collection protocols. POET records five observers viewing each image for one second, while COCO-Search18 records ten observers performing an explicit search task. The closer correspondence between Sparsity in Table 1 and KL divergence in Table 2 on POET indicates that concentrated explanations tend to receive larger distributional penalties against its sparser fixation observations. This association should not be interpreted causally because the datasets also differ in images, categories, observer counts, and viewing tasks. The gaze results therefore measure agreement with the recorded behavioral distributions rather than the universal superiority of one attribution profile.

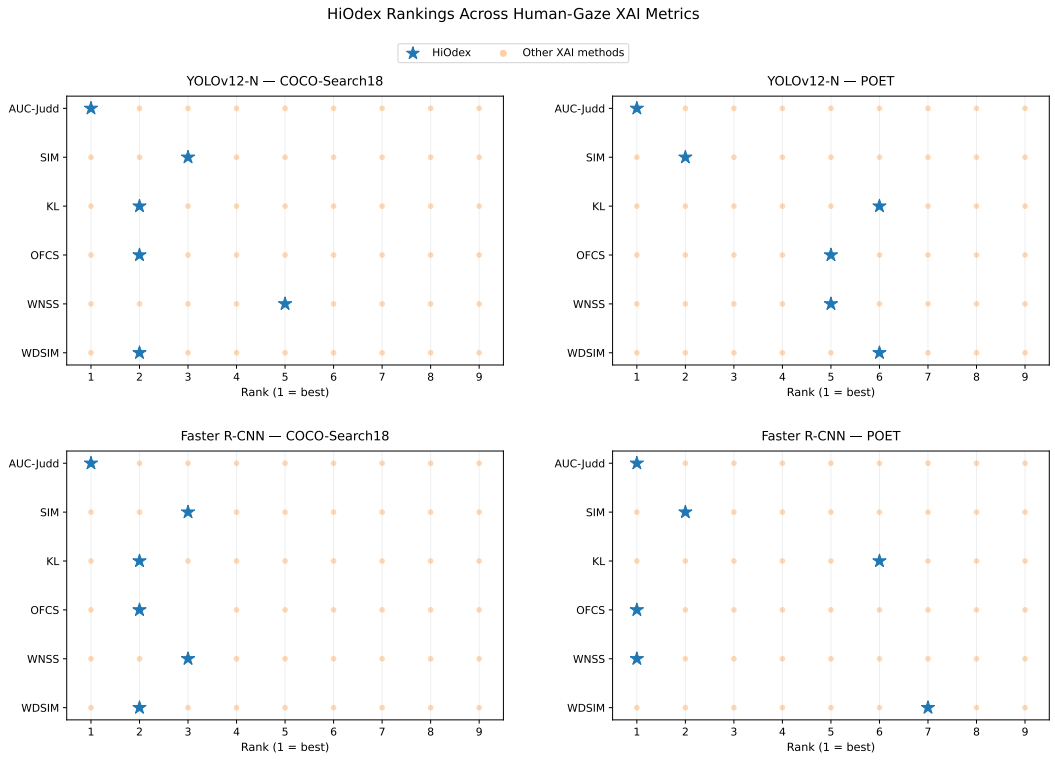

Figure C1: Method ranks across the six human-gaze metrics. Stars denote HiOdex, circles denote the eight baselines, and rank 1 is best.

Computational trade-off. The runtimes plotted in Fig. 1 follow the matched timing protocol in App. B.7. The fixation-count analysis in App. C.4 further shows that additional fixations increase runtime without producing a corresponding improvement in the model-oriented metrics.

C.2 Ablation Protocol

Each principal HiOdex parameter is varied independently while all remaining parameters are held at their reference values. Table C1 specifies the reference configuration and complete evaluation grid. The same settings are applied across all four detector and benchmark combinations, using the same images and detections. The condition $\rho = \text{off}$ removes the object-context gate. The fusion-weight sweep is evaluated only with YOLOv12-N because its P3, P4, and P5 layers form an ordered feature hierarchy, whereas the selected Faster R-CNN feature-pyramid layers use uniform fusion.

Table C1: Hyperparameter settings used in the ablation study.

Parameter	**Reference**	**Evaluation values**
δ	0.40	$\{0.05, 0.10, 0.20, 0.40, 0.60, 1.00\}$
K	3	$\{1, 2, 3, 5, 8, 10\}$
γ	0.04	$\{0.00, 0.01, 0.02, 0.04, 0.06\}$
β	0.50	$\{0.0, 0.2, 0.4, 0.5, 0.6, 0.8, 1.0\}$
ρ	0.60	$\{0.2, 0.4, 0.6, 0.8, 1.0, \text{off}\}$
$\boldsymbol{\omega}$	$(0.55, 0.35, 0.10)$	$\{(0.55, 0.35, 0.10), (1, 0, 0), (0, 1, 0),$ $(0, 0, 1), (0.33, 0.33, 0.33)\}$

Every parameter sweep reports PG, EBPG, OA, and Sparsity, averaged over the fixed evaluation samples within each setting, for which higher values indicate better performance. Deletion and Insertion are evaluated at every setting and summarized by OA, which measures their difference. Attribution-generation time is additionally reported for the number of fixations K, using the timing protocol in App. B.7. These experiments assess model-oriented explanation behavior and do not estimate how the method parameters affect human-gaze alignment.

C.3 Summary of Mechanism Ablations

This subsection asks which design choices account for HiOdex's behavior. We isolate four mechanisms: sequential fixation selection, complementary attribution branches, object-context gating, and multi-layer fusion. Table C2 states the controlled change applied to each mechanism, the role it is expected to serve, and the response observed across the evaluated settings.

Fig. C2 complements the table at the parameter level. It reports the direction in which each metric responds as each of the six swept parameters increases, including the fixation-window width δ and the attribution-smoothing width γ, which the four mechanisms do not cover. For $\boldsymbol{\omega}$, the direction denotes shifting fusion weight from P3 toward P5 and applies only to the two YOLOv12-N settings. The fixation-window width has limited effect across its evaluation range, whereas increasing the smoothing width trades attribution concentration for peak reliability.

The fixation process reaches diminishing returns quickly. Most of the explanation structure is established by the first two or three selections, after which the quality metrics remain nearly unchanged while computation continues to increase. We therefore use $K = 3$ as a stable trade-off between evidence coverage and efficiency. The branch comparison reveals an asymmetric relationship. Removing the Local Evidence Map consistently weakens performance, whereas retaining it alone remains competitive under the model-oriented metrics. Intermediate mixtures provide the most stable behavior, indicating that fixation-guided context should complement rather than replace position-wise local evidence.

The gate and fusion interventions produce stronger responses. Removing the object-context gate consistently weakens localization and concentration, while the broad OA maximum near $\rho = 0.6$ favors retaining nearby context over either a restrictive object mask or unrestricted spatial support. Feature resolution is similarly important. Configurations retaining substantial P3 contribution remain competitive, whereas restricting attribution to P4 or P5 progressively removes spatial

Table C2: Controlled ablations of the principal HiOdex mechanisms. Higher is better for PG, EBPG, OA, and Sparsity. Fusion is evaluated only with YOLOv12-N. Detailed sweeps are reported in Appendices C.4–C.6.

Mechanism	Controlled change	Expected role	Observed response
Fixation process	$K \in \{1, 2, 3, 5, 8, 10\}$; reference $K = 3$	Successive fixations extend coverage to distinct evidence-bearing regions.	PG, EBPG, OA, and Sparsity change little after the first few fixations, whereas runtime increases steadily.
Branch integration	$\beta = 0$ retains only the Local Evidence Map, and $\beta = 1$ retains only the Foveal CAM; reference $\beta = 0.5$	The Foveal CAM supplies fixation-guided context, while the Local Evidence Map preserves position-wise detail.	Intermediate mixtures remain stable. The Local Evidence Map alone remains competitive, whereas the Foveal CAM alone is the weakest configuration.
Object-context gate	$\rho \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ and gate removal; reference $\rho = 0.6$	The gate suppresses distant responses while retaining context around the selected object.	OA reaches a broad interior maximum near 0.6. Removing the gate produces the lowest PG, EBPG, and Sparsity in all four settings.
Multi-layer fusion	Reference, uniform, P3-only, P4-only, and P5-only fusion	Fusion combines the spatial precision of fine features with the contextual selectivity of deeper features.	The reference, uniform, and P3-only settings remain competitive. P4-only fusion causes moderate degradation, while P5-only fusion substantially degrades every principal metric.

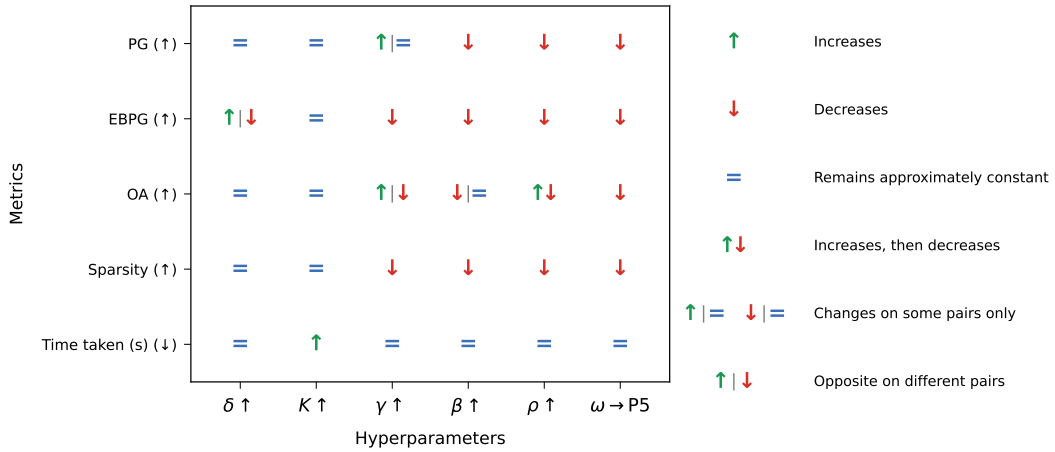

Figure C2: Directions of the metric responses across the HiOdex parameter sweeps.

detail. Because P3-only fusion remains competitive, this experiment supports the importance of high-resolution evidence more strongly than the necessity of multi-layer fusion itself. The deeper feature levels appear to provide complementary context rather than the primary localization signal. These patterns support a compact and stable operating configuration without implying that every reference value is uniquely optimal.

C.4 Fixation Process

Fixation-window width Fig. C3 and C4 jointly characterize the quantitative and qualitative responses to the fixation-window width δ. PG, OA, and Sparsity remain approximately constant throughout most of the evaluated range and have largely stabilized by $\delta = 0.2$ to 0.4. EBPG is more sensitive, increasing with YOLOv12-N on MS-COCO while decreasing in the other three settings. The attribution maps likewise change only modestly after this interval. A moderate window is therefore sufficient without detector-specific or instance-specific adjustment, and $\delta = 0.4$ is used throughout the main evaluation.

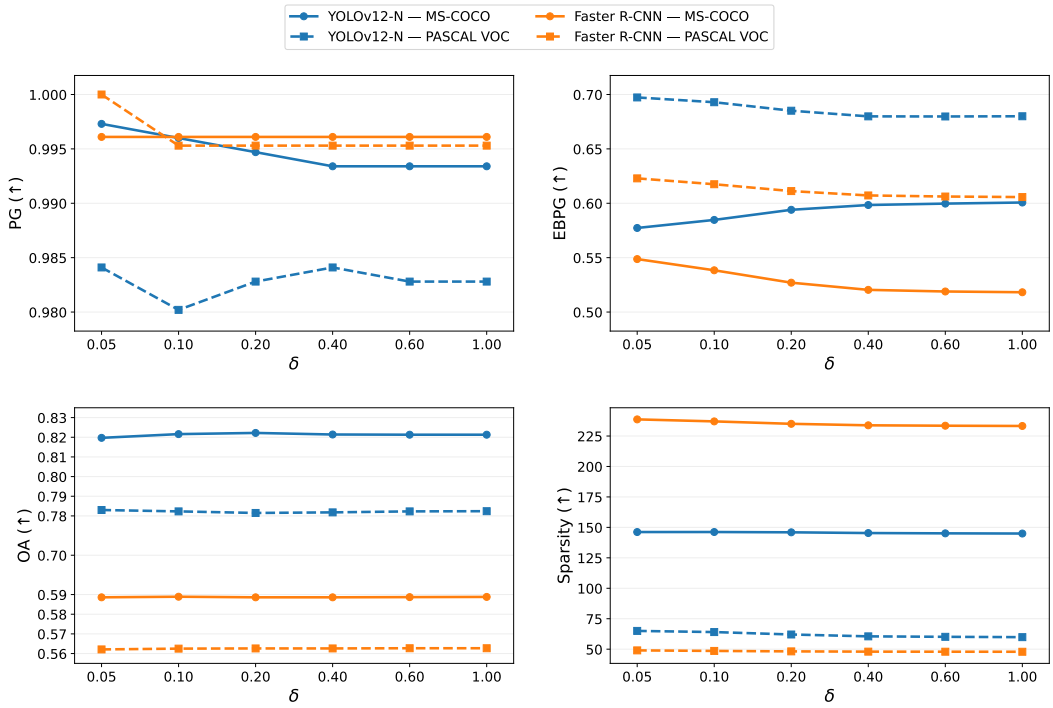

Figure C3: Sensitivity to the fixation-window width δ across both detectors and datasets.

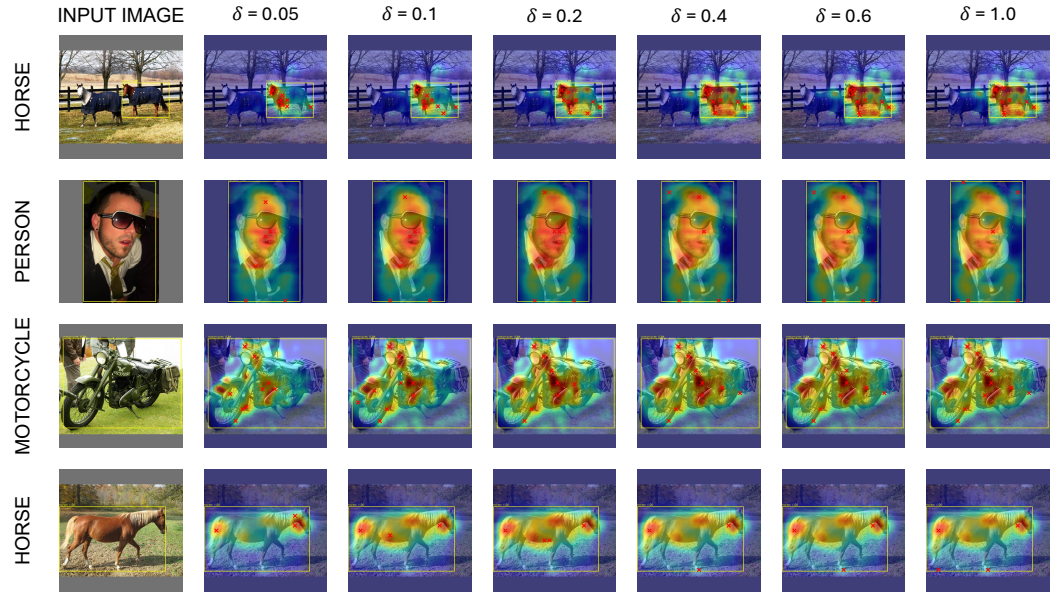

Figure C4: Attribution maps across the fixation-window sweep. Yellow boxes indicate selected detections, and red crosses indicate model-generated fixations.

Number of fixations Fig. C5 and C6 show that PG, EBPG, OA, and Sparsity are near saturation at the smallest evaluated values of K and remain essentially unchanged through $K = 10$. The principal attribution structure is also established after the first few fixations. Attribution-generation

time, however, increases steadily with K. Additional fixations therefore add computation without a corresponding improvement under these criteria. We use $K = 3$ to preserve the sequential fixation mechanism while avoiding unnecessary computation.

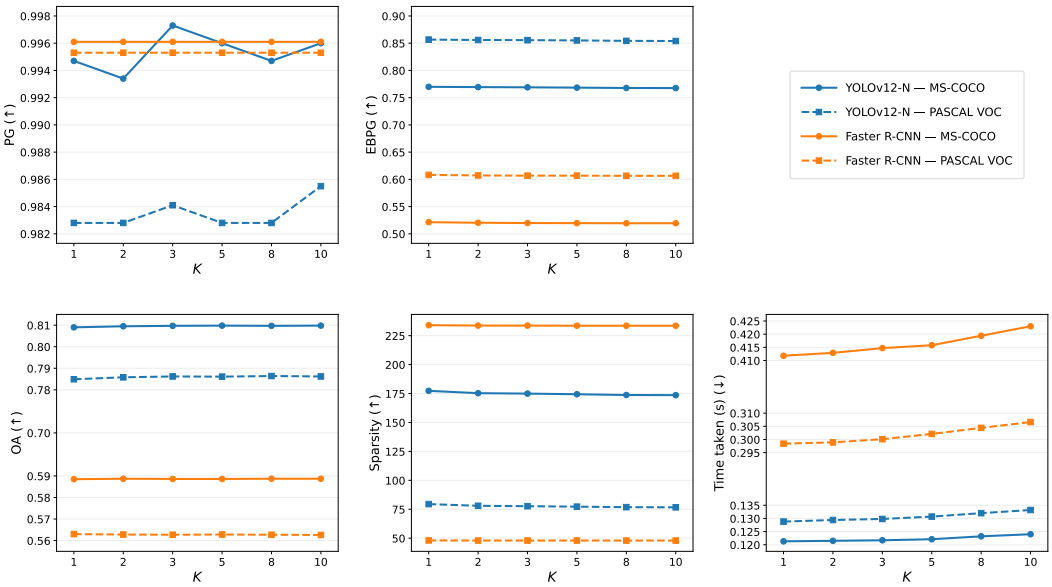

Figure C5: Sensitivity to the number of fixations K, including attribution-generation time.

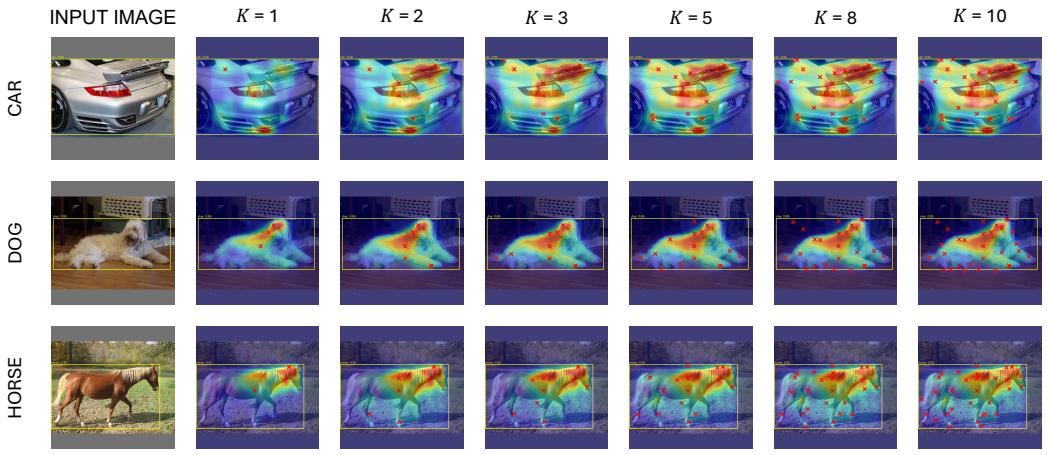

Figure C6: Attribution maps across the fixation-count sweep. Yellow boxes indicate selected detections, and red crosses indicate model-generated fixations.

C.5 Attribution Construction and Object Context

Attribution-smoothing width Fig. C7 and C8 show that increasing γ reduces EBPG and Sparsity in all four settings, indicating that smoothing broadens the attribution distribution and reduces extreme concentration. It improves PG under YOLOv12-N, where peak localization has not already saturated. On MS-COCO, PG reaches its maximum and stabilizes near $\gamma = 0.04$, while on PASCAL VOC it continues to increase through $\gamma = 0.06$. Faster R-CNN changes less because its PG

scores are already close to saturation. The qualitative examples show the corresponding suppression of fine-scale variation while preserving the principal object-focused regions. We use $\gamma = 0.04$ as a balanced operating point between peak reliability and attribution concentration.

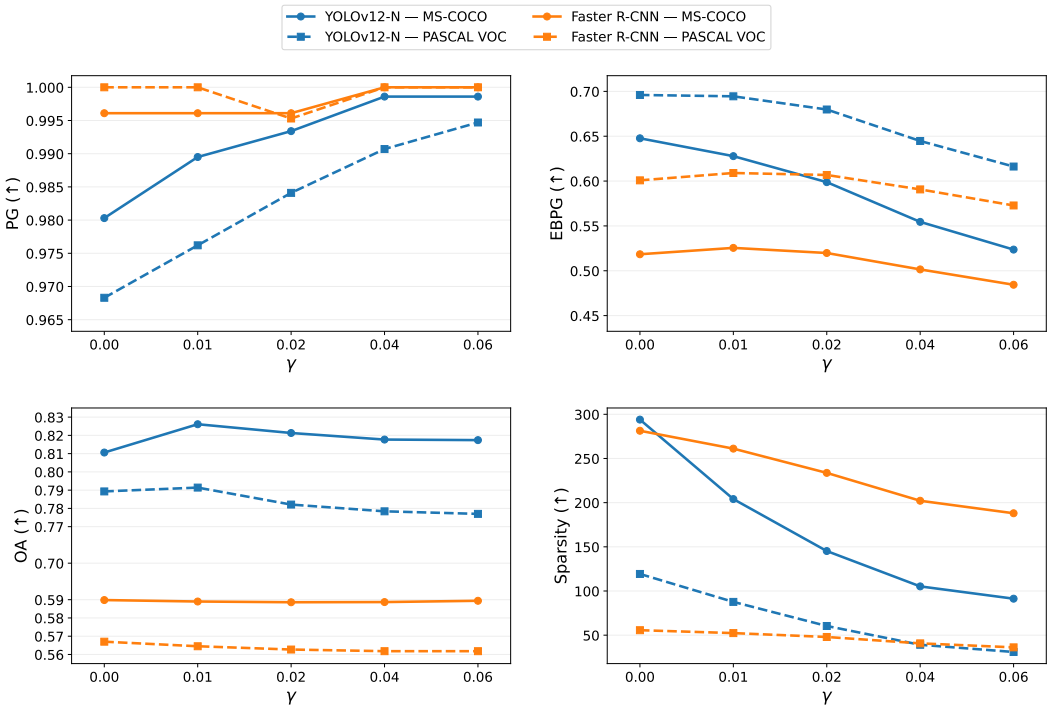

Figure C7: Sensitivity to the attribution-smoothing width γ across both detectors and datasets.

Branch-mixture weight The endpoints $\beta = 0$ and $\beta = 1$ isolate the Local Evidence Map and Foveal CAM branches, respectively. Fig. C9 and C10 show that intermediate values, particularly $\beta \in [0.4, 0.6]$, remain stable across both detectors and datasets. Increasing β toward one consistently reduces EBPG and generally lowers Sparsity, eventually degrading PG and OA. The qualitative sequence similarly reveals the transition from localized evidence to the broader Foveal CAM response. These results show that the Foveal CAM alone is insufficient. However, the model-oriented metrics do not establish that an intermediate mixture outperforms the Local Evidence Map alone. We retain $\beta = 0.5$ because it preserves both attribution branches within the stable region, not because it represents a uniquely optimal value.

Object-context gate Fig. C11 and C12 show that OA follows a broad rise-and-fall profile across the continuous values of ρ, with $\rho = 0.6$ lying near the maximum under both detectors and datasets. Increasing ρ generally reduces EBPG and Sparsity and produces a smaller decline in PG. More importantly, removing the gate yields the lowest PG, EBPG, and Sparsity in every setting. The corresponding attribution maps illustrate the loss of object-conditioned spatial support when the gate is removed. This controlled comparison shows that the gate contributes to localization and attribution concentration. We use $\rho = 0.6$ as a stable compromise among faithfulness, localization, and concentration.

C.6 Multi-Layer Fusion

Fig. C13 and C14 compare the reference fusion weights with uniform fusion and the three single-level alternatives under YOLOv12-N. The reference configuration, uniform fusion, and P3-only attribution remain closely comparable because each preserves substantial high-resolution information. Restricting attribution to P4 moderately degrades all four criteria, while using only P5 causes a

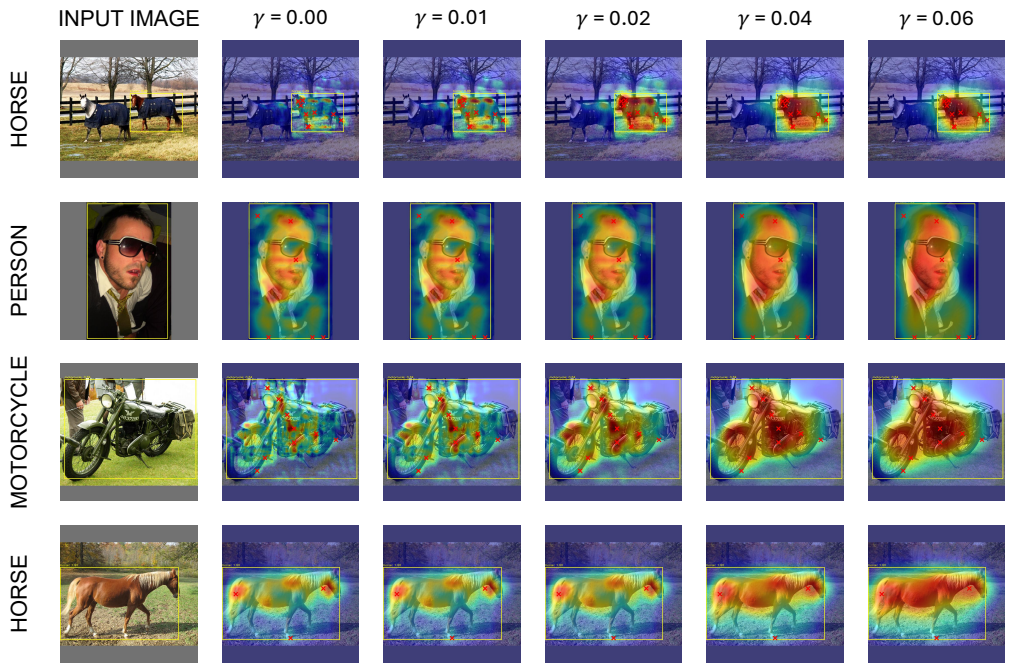

Figure C8: Attribution maps across the smoothing-width sweep. Yellow boxes indicate selected detections, and red crosses indicate model-generated fixations.

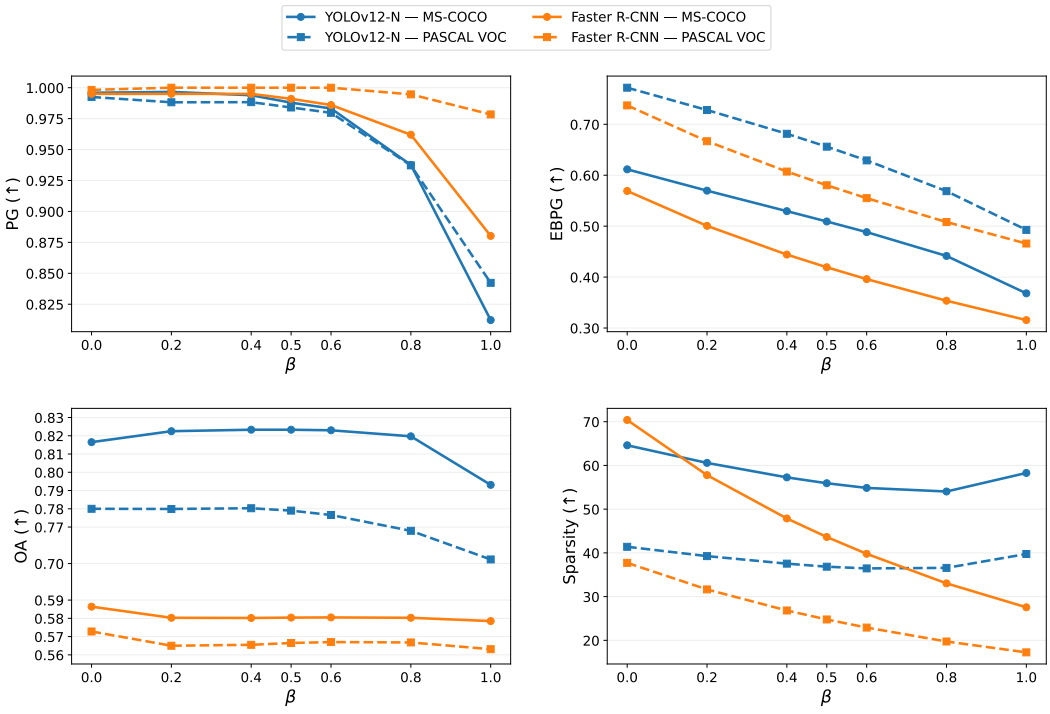

Figure C9: Sensitivity to the branch-mixture weight β across both detectors and datasets.

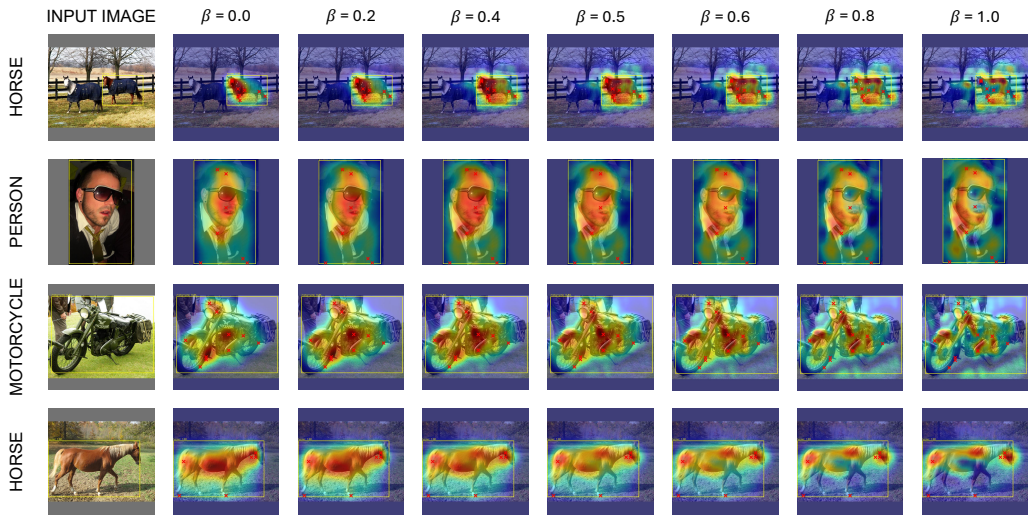

Figure C10: Attribution maps across the branch-mixture sweep. Yellow boxes indicate selected detections, and red crosses indicate model-generated fixations.

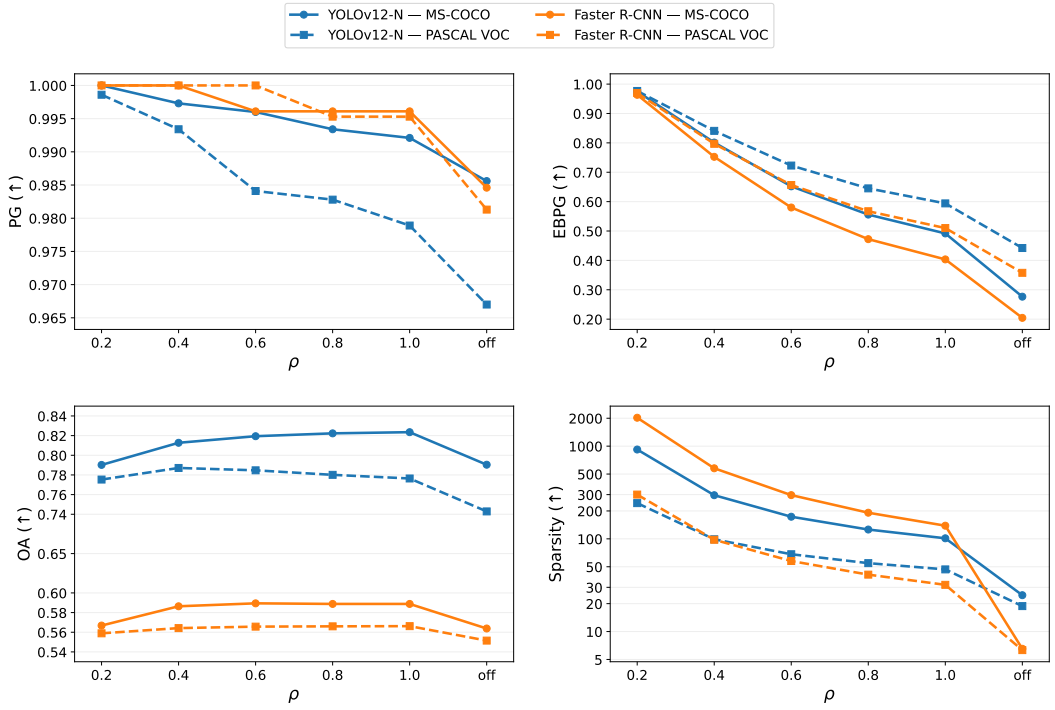

Figure C11: Sensitivity to the object-context gate width ρ, including the gate-removed condition.

pronounced collapse. The degradation is smaller on PASCAL VOC than on MS-COCO but follows the same direction. The attribution maps show a corresponding loss of spatial detail as fusion shifts toward coarser representations. Multi-layer fusion therefore depends more strongly on retaining fine-resolution evidence than on the precise weights assigned among configurations that preserve P3.

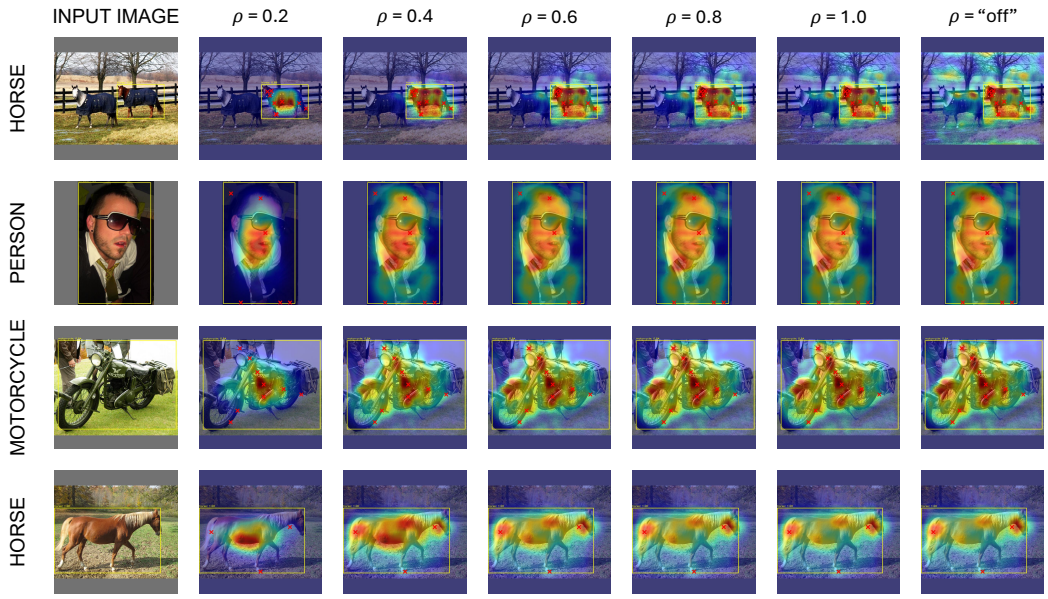

Figure C12: Attribution maps across the object-context-gate sweep, including the gate-removed condition. Yellow boxes indicate selected detections, and red crosses indicate model-generated fixations.

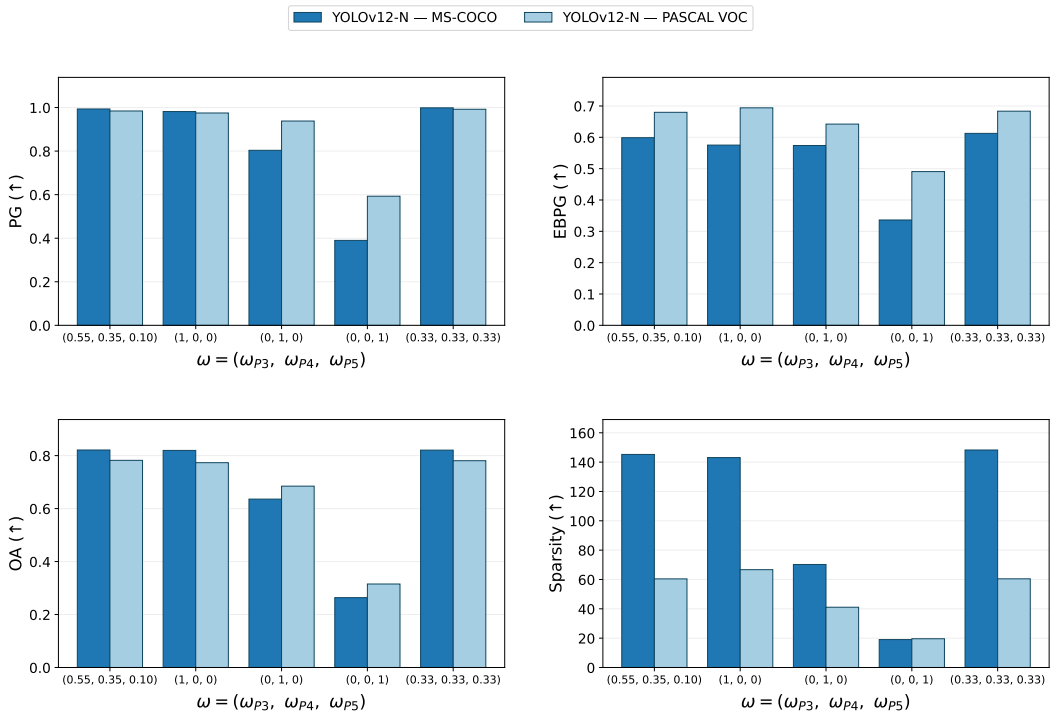

Figure C13: Sensitivity to the YOLOv12-N multi-layer fusion weights ω.

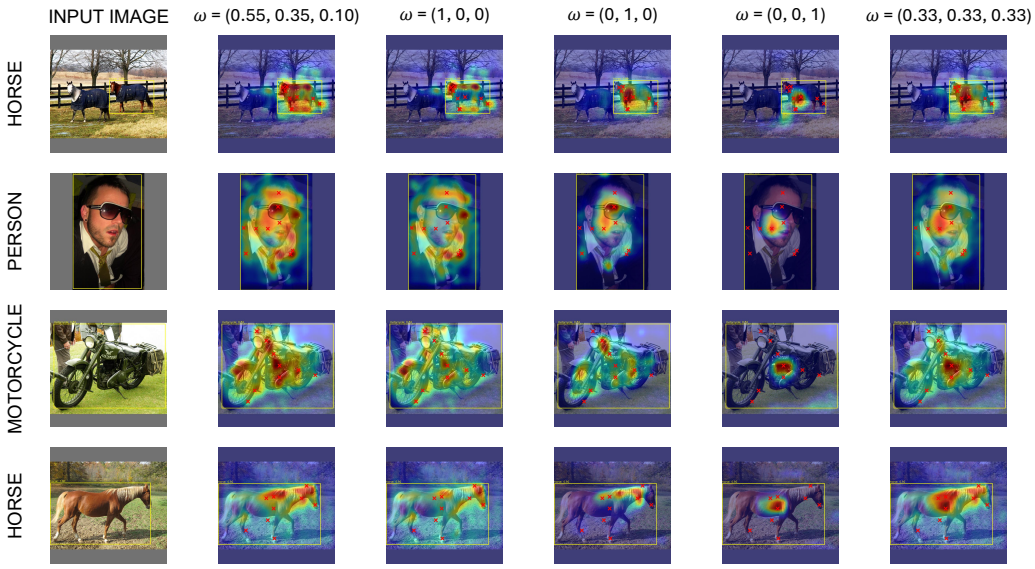

Figure C14: Attribution maps under the evaluated multi-layer fusion policies. Yellow boxes indicate selected detections, and red crosses indicate model-generated fixations.

C.7 Gaze-Metric Sensitivity

Fig. C15 examines how the parameter-dependent gaze metrics respond to their clustering and density-estimation settings. For OFCS, WNSS, and WDSIM, the DBSCAN radius is varied over $\mathrm{Eps} \in \{1°, 2°\}$, and MinPts is varied over $\{1, 3, 5\}$. For SIM and KL divergence, the fixation-density bandwidth is varied over $\sigma \in \{0.5°, 1.0°, 1.5°, 2.0°\}$. Within each detector and gaze-dataset setting, every result is averaged across the nine explanation methods. AUC-Judd does not depend on these parameters and is therefore excluded from the sweep. OFCS, WNSS, and WDSIM generally increase with MinPts, while increasing σ generally raises SIM and lowers KL divergence. No configuration is uniformly optimal across the five parameter-dependent metrics and four evaluation settings. We therefore use $\mathrm{Eps} = 1°$, $\mathrm{MinPts} = 3$, and $\sigma = 1°$ as fixed reference settings rather than universal optima.

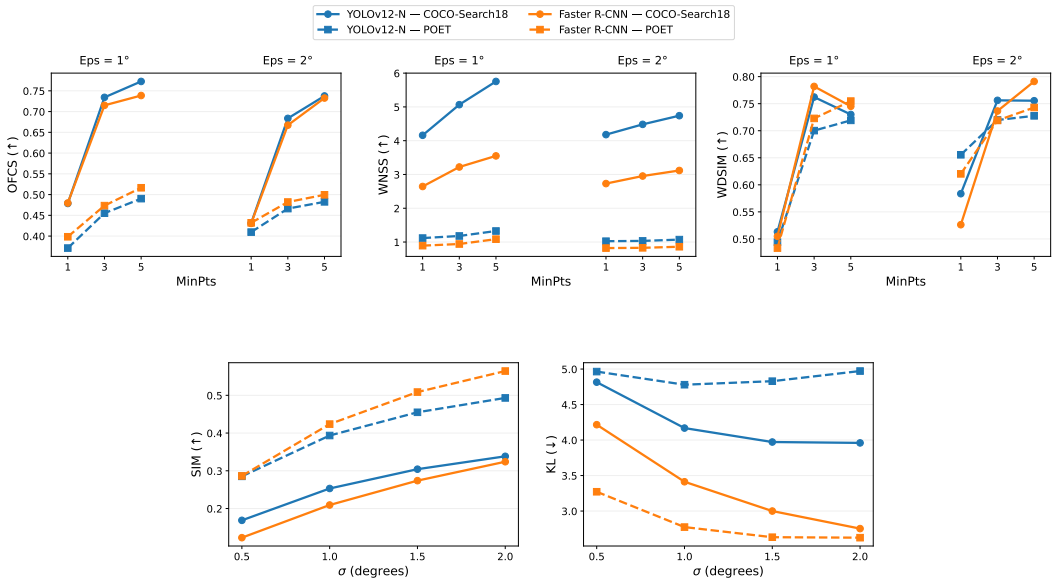

Figure C15: Sensitivity of the parameter-dependent gaze metrics to Eps, MinPts, and σ.

D Additional Evaluation and Qualitative Evidence

This appendix evaluates HiOdex on specialized industrial imagery and provides additional cross-method attribution comparisons. App. D.1 examines whether the principal model-oriented results extend to the Substation Equipment dataset. App. D.2 presents complementary YOLOv12-N and Faster R-CNN examples that are not required for the main argument but help characterize method-specific attribution patterns.

D.1 Evaluation on the Substation Equipment Dataset

The main evaluation uses MS-COCO and PASCAL VOC as general-purpose object-detection benchmarks. We additionally apply the same evaluation protocol to the Substation Equipment dataset (Gomes et al., 2023), which contains many small, visually similar objects embedded in complex industrial structures. The experiment retains the nine explanation methods, two detector architectures, and six model-oriented metrics introduced in Sec. 4.

The dataset contains 1,660 electrical-substation images annotated across 15 equipment classes. Originally released for semantic segmentation, it is reorganized for object detection into 1,142 training, 244 validation, and 274 test images, yielding approximately 50,596 object annotations after duplicate removal. We use this dataset as a complementary domain-specific evaluation rather than an established detection benchmark. Its small targets, limited inter-class variation, and structural clutter test whether attribution remains associated with the selected detection rather than nearby equipment or background structure.

Table D1 reports the complete quantitative comparison. HiOdex achieves the highest OA under both detector architectures. With YOLOv12-N, it obtains 0.9181, narrowly exceeding D-RISE at 0.9167 and outperforming D-CLOSE at 0.8779. With Faster R-CNN, HiOdex reaches 0.7782, compared with 0.7720 for D-RISE and 0.7507 for D-CLOSE. It also achieves the highest Insertion score in both settings.

No method dominates every criterion. G-CAME achieves the highest PG, EBPG, and Sparsity under both detectors, while D-CLOSE obtains the lowest Deletion score. This result reinforces the distinction established in Sec. 5.1: concentrated attribution within a predicted object does not necessarily provide the strongest combined Insertion and Deletion behavior. HiOdex instead preserves the OA pattern observed on the general-purpose benchmarks.

Table D1: Model-oriented explanation results on the Substation Equipment dataset. Best values are shown in bold, and second-best values are underlined. † denotes human-inspired XAI methods.

		Faithfulness			Localization		
Detector	Method	OA ↑	Del. ↓	Ins. ↑	PG ↑	EBPG ↑	Spars. ↑
YOLOv12-N	HiOdex (Ours)†	**.9181**	.0022	**.9203**	.8534	.1374	106.9250
	FovEx†	.8553	.0318	.8871	.0495	.0143	6.4796
	G-CAME	.8555	.0022	.8578	**.9742**	**.6217**	**2260.5617**
	Grad-CAM	.7852	.0191	.8043	.1683	.0334	18.6881
	Grad-CAM++	.7645	.0273	.7918	.1241	.0204	11.0907
	Eigen-CAM	.7214	.0475	.7690	.0382	.0105	6.0216
	LayerCAM	.5593	.0144	.5738	.9557	.5364	1106.5233
	D-RISE	.9167	.0023	.9190	.5940	.0159	7.4994
	D-CLOSE	.8779	**.0017**	.8797	.9213	.0560	48.4268
Faster R-CNN	HiOdex (Ours)†	**.7782**	.0022	**.7804**	.9545	.1131	60.9068
	FovEx†	.7108	.0243	.7351	.0593	.0213	6.6132
	G-CAME	.7552	.0022	.7574	**.9851**	**.4060**	**651.0169**
	Grad-CAM	.3996	.0551	.4548	.0732	.0094	3.9377
	Grad-CAM++	.1534	.1605	.3139	.0188	.0081	3.4864
	Eigen-CAM	-.1353	.3862	.2508	.0029	.0062	2.8239
	LayerCAM	.6330	.0019	.6349	.7645	.2765	222.7278
	D-RISE	.7720	.0017	.7737	.5243	.0223	8.4930
	D-CLOSE	.7507	**.0015**	.7521	.7135	.0435	23.8047

Fig. D1 complements the quantitative comparison with aligned examples from both detectors. In the displayed cases, HiOdex remains associated with the selected equipment instance despite nearby objects and structural clutter. The baselines range from sharply concentrated responses to broader activation over surrounding structures. Together, Table D1 and Fig. D1 show that the principal OA pattern extends to this specialized small-object setting. This experiment does not establish general robustness across industrial domains.

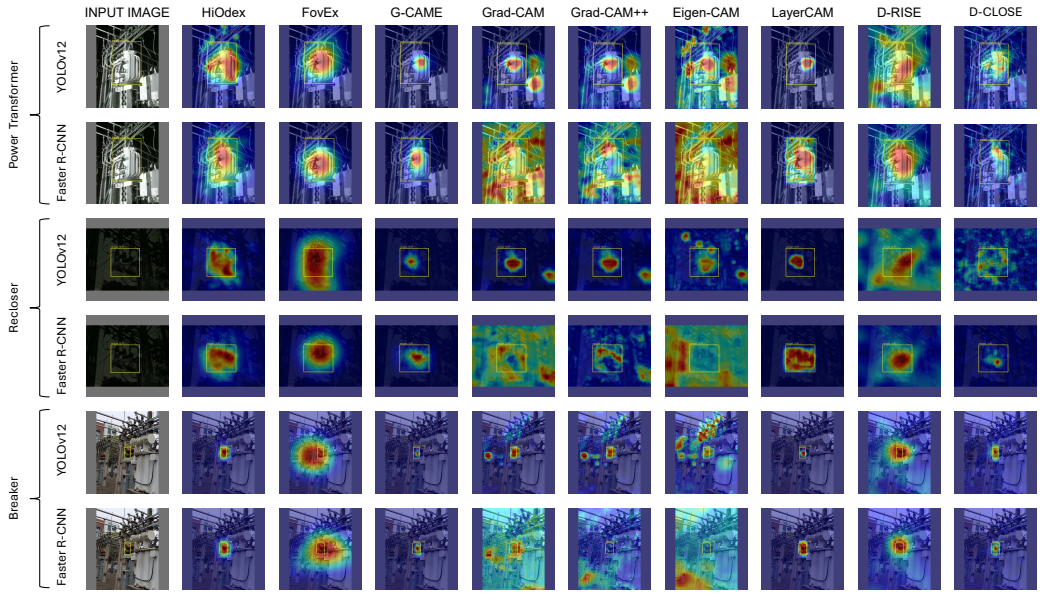

Figure D1: Cross-method attribution comparison on the Substation Equipment dataset. Each pair of rows shows one selected detection under YOLOv12-N and Faster R-CNN, and columns compare HiOdex with the eight baselines. Yellow boxes indicate the selected detections, and red crosses denote the model-generated fixations used by HiOdex and FovEx.

D.2 Additional Cross-Method Attribution Comparisons

The following panels extend the qualitative comparison beyond the examples presented in Sec. 5.1. They cover variation in object scale, aspect ratio, scene context, and the number of nearby instances. Each row corresponds to one selected detection, with identical visualization conventions across methods. These panels are diagnostic rather than quantitative. The comparative conclusions remain grounded in the quantitative results in Sec. 5 and the ranking analysis in App. C.1.

Fig. D2 presents complementary YOLOv12-N examples not included in the main benchmarking panel. In the displayed cases, HiOdex remains associated with the selected detection while distributing attribution across multiple supporting regions within or near its predicted box. G-CAME and LayerCAM often emphasize compact discriminative regions, whereas several Grad-CAM-family and Eigen-CAM explanations extend more broadly into the surrounding scene. FovEx and D-RISE generally produce smoother responses, while D-CLOSE distributes finer attribution across the selected region. These differences illustrate why localization, concentration, and perturbation-based faithfulness can rank the same explanation differently.

Fig. D3 provides the corresponding complete comparison for Faster R-CNN. HiOdex remains associated with the selected detection across substantial variation in object size, shape, and surrounding context. The locations and spatial extent of several baseline responses differ from their YOLOv12-N counterparts, ranging from sharply localized evidence to broader or fragmented scene-level patterns. These examples demonstrate that attribution structure can depend on detector architecture and complement the cross-detector quantitative evaluation in Sec. 5.

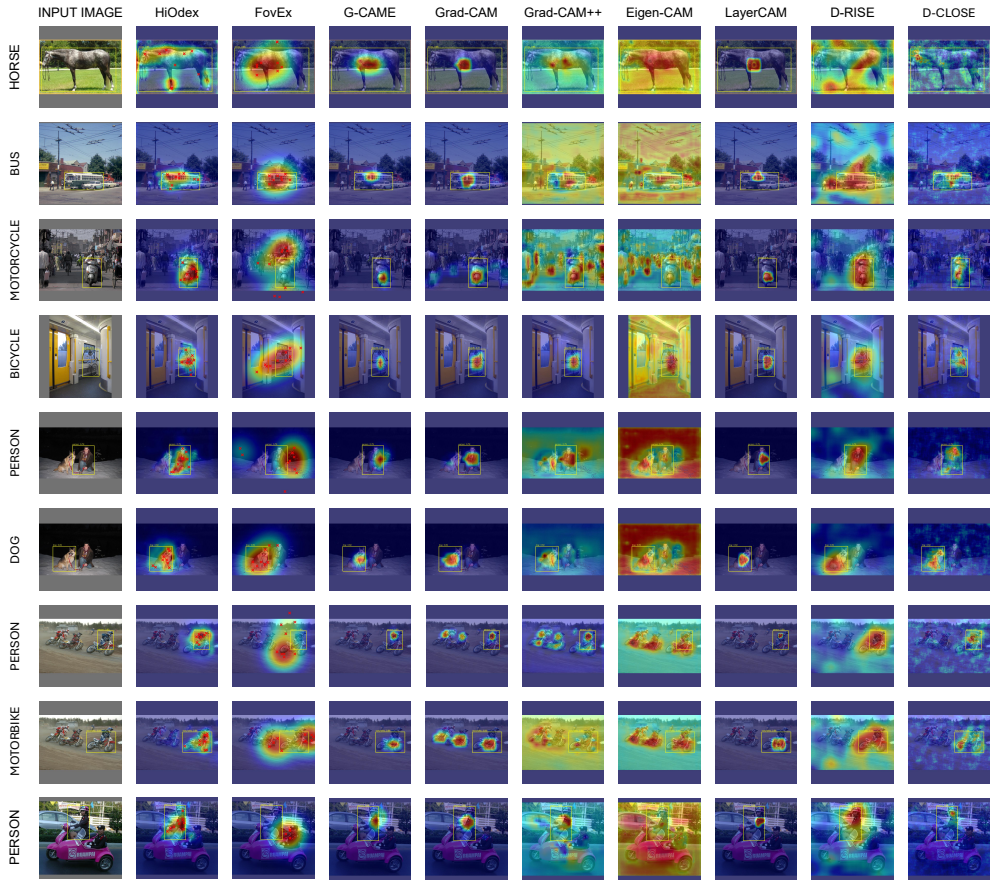

Figure D2: Additional cross-method attribution comparisons with YOLOv12-N. Rows correspond to selected detections, and columns compare HiOdex with the eight baselines. Yellow boxes indicate the selected detections, and red crosses denote the model-generated fixations used by HiOdex and FovEx.

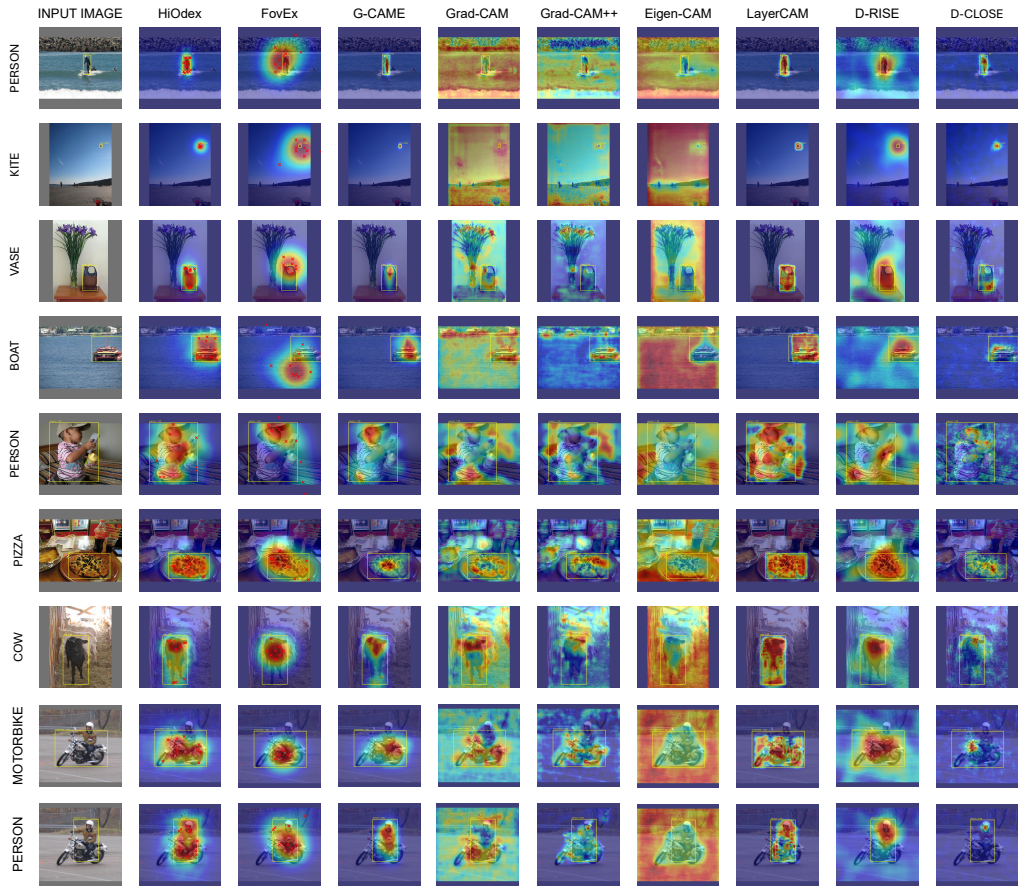

Figure D3: Additional cross-method attribution comparisons with Faster R-CNN. Rows correspond to selected detections, and columns compare HiOdex with the eight baselines. Yellow boxes indicate the selected detections, and red crosses denote the model-generated fixations used by HiOdex and FovEx.