

Explainable and Uncertainty Aware Deep Learning for Concealed Prohibited Item Detection in Nonintrusive X-ray Inspection

Paul Kategaya Kisozi

paulkisozi61@gmail.com

Cavendish University Uganda

Research Article

Keywords: prohibited-item detection, concealment-aware learning, explainable AI, uncertainty estimation, YOLO26

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10974557/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Explainable and Uncertainty Aware Deep Learning for Concealed Prohibited Item Detection in Nonintrusive X-ray Inspection

Paul Kategaya Kisozi*

*¹Department of Information, Communication and Technology, Cavendish
University Uganda, Kampala, Uganda*

***Corresponding Author:** Paul Kategaya Kisozi; *Email:* paulkisozi61@gmail.com

Abstract

Prohibited-item detection in security X-ray images remains challenging because objects may be small, overlapping, weakly contrasted, or deliberately concealed within complex baggage contents. This study developed and evaluated CAU-XInspect, a concealment-aware, explainable, and uncertainty-aware deep learning framework for prohibited-item X-ray detection. The framework was built on a YOLO26 detector and introduced a Concealment-Aware Feature Enhancement Module (CAFEM) that integrates occlusion-sensitive, contrast-sensitive, and contextual feature refinement, together with mask guidance and attention-mask consistency learning. Experiments were conducted on the PIDray benchmark comprising 47,677 X-ray images across 12 prohibited-item classes, with 29,457 images used for development and 18,220 held-out Easy, Hard, and Hidden test images. A controlled B1–B6 ablation study was conducted to quantify the incremental effects of conventional attention (CBAM), concealment-aware attention (CAAM), the proposed CAFEM, mask guidance, and attention-mask consistency learning. The B5 configuration achieved the highest overall mAP@0.50:0.95 of 0.5991 and mAP@0.50 of 0.7435, while the full B6 model achieved the highest overall recall of 0.7111 and the strongest concealed-item sensitivity. On the Hidden subset, B6 improved recall from 0.3550 for the YOLO26 baseline to 0.4918, representing a relative improvement of approximately 38.5%, while Hidden F1-score increased from 0.4600 to 0.5478. Spatial explainability also improved progressively, with B6 achieving an attention IoU of 0.7578 and mask IoU of 0.7777. Reliability analysis showed low calibration error under the validation protocol (ECE 0.0095, Brier score 0.0146) and strong error discrimination using detector confidence (AUC 0.9894) and Monte Carlo dropout uncertainty (AUC 0.8873). However, zero-shot evaluation on an independent 831-image SIXray-derived subset produced a substantially lower mAP@0.50:0.95 of 0.0408 and recall of 0.0954, demonstrating considerable cross-dataset domain shift. Collectively, these findings show that concealment-aware feature enhancement can improve sensitivity and spatial grounding for difficult prohibited-item X-ray cases while providing complementary reliability information. Nevertheless, CAU-XInspect should be regarded as a technical proof of concept rather than an operationally validated security-screening system. Broader cross-scanner validation, domain adaptation, and human-in-the-loop evaluation are required before real-world deployment.

Keywords: prohibited-item detection, concealment-aware learning, explainable AI, uncertainty estimation, YOLO26.

Article Highlights

- i. Introduces CAU-XInspect, a concealment-aware, explainable, and uncertainty-aware security X-ray detection framework that combines YOLO26 with CAFEM, mask guidance, and attention–mask consistency learning.
- ii. Improves concealed-item sensitivity, with the full B6 model increasing Hidden recall from 0.3550 to 0.4918 and Hidden F1-score from 0.4600 to 0.5478, while also achieving stronger spatial grounding through attention and mask IoU.
- iii. Integrates detection, explainability, calibration, uncertainty, and external validation, showing strong internal PIDray reliability while revealing substantial zero-shot domain shift on an independent SIXray-derived test set.

1 Introduction

Non-intrusive inspection (NII) is a central component of contemporary security screening because baggage and cargo must be examined for prohibited or dangerous items without physically opening every container [1],[2]. X-ray imaging supports this process by revealing internal object structure, material density, and overlap patterns that are not visible from the exterior [3],[4]. However, automated interpretation remains difficult because security X-ray scenes frequently contain dense clutter, overlapping benign objects, weak object boundaries, and partially concealed prohibited items[5],[6][3]. These difficulties are operationally important because missed prohibited items can create substantial security risk, whereas excessive false alarms can increase manual inspection burden and delay screening[7].

Recent advances in artificial intelligence (AI) and deep learning have substantially improved automated prohibited-item detection. Large-scale X-ray benchmarks such as SIXray, OPIXray, and PIDray have enabled the development and evaluation of modern detection architectures under increasingly realistic conditions[8],[9],[10]. SIXray emphasized extreme class imbalance, overlapping objects, and sparse prohibited-item occurrence [10]. OPIXray focused more explicitly on occlusion and introduced de-occlusion attention for partially visible objects [9]. PIDray extended the problem further by providing large-scale real-world X-ray images, bounding-box and segmentation information, deliberately hidden prohibited items, and Easy, Hard, and Hidden test subsets [8]. These datasets have encouraged a shift from generic object detection toward X-ray-specific representation learning using attention, context modelling, feature aggregation, and foreground-background separation.

Despite these advances, strong aggregate detection performance does not fully resolve the requirements of safety-critical X-ray inspection [2]. First, concealment and object overlap can suppress the visual evidence needed for reliable detection, particularly on Hidden and highly cluttered cases [11],[12],[13]. Second, many high-performing detectors provide only a bounding box, class label, and confidence score, with limited evidence that the highlighted regions

correspond to the actual prohibited object[14],[15]. Third, neural-network confidence is not equivalent to reliability as incorrect detections may still receive high confidence, which creates a need for calibration and uncertainty estimation[16]. Finally, a model that performs strongly on one benchmark may degrade substantially when applied to another scanner, pseudocoloring scheme, acquisition environment, or label distribution[17]. Therefore, detection accuracy, concealment sensitivity, spatial explainability, predictive reliability, and cross-dataset robustness should be considered together rather than as isolated objectives.

Existing research has addressed important parts of this problem, but usually in separate methodological streams. De-occlusion attention has been used to recover partially visible items [9], selective dense attention has been applied to hidden-object detection in PIDray [8], foreground-background separation has been proposed to suppress irrelevant baggage clutter [5] and global context-aware multi-scale aggregation has been used to strengthen recognition of small or occluded prohibited items [7]. Other studies have investigated domain discrepancy across X-ray datasets[18],[17]. However, fewer studies integrate concealment-aware feature refinement with quantitative explanation evaluation, calibrated uncertainty, controlled component-wise ablation, and independent cross-dataset testing within one prohibited-item inspection framework.

This study addresses that gap through Concealment-Aware Uncertainty-aware Explainable X-ray Inspection (CAU-XInspect), a concealment-aware, explainable, and uncertainty-aware X-ray inspection framework built around a strong YOLO26 baseline. The proposed method introduces a Concealment-Aware Feature Enhancement Module (CAFEM) that combines occlusion, contrast, and contextual cues immediately before the detection head. The framework progressively evaluates conventional attention, concealment-aware attention, CAFEM, mask guidance, and consistency learning under a controlled B1-B6 ablation protocol. It further evaluates whether the learned spatial evidence aligns with prohibited-item regions, whether calibration and Monte Carlo dropout uncertainty discriminate less reliable predictions, and whether the final PIDray-trained model transfers without fine-tuning to a SIXray-derived external test set. The emphasis is therefore not simply on obtaining the largest possible mean Average Precision (mAP) value, but on determining whether a strong detector can become more sensitive to concealed items, more spatially grounded, and more reliability-aware without losing the practical advantages of modern one-stage detection.

The main contributions of the study are fivefold:

- i. Concealment-aware feature enhancement: CAFEM was developed to combine occlusion, contrast, and contextual information before the detection head, with mask guidance and consistency learning evaluated as progressive extensions.
- ii. Controlled component-wise evaluation: a B1-B6 ablation sequence was used, beginning with YOLO26 and progressively adding Convolutional Block Attention Module (CBAM), Concealment-Aware Attention Module (CAAM), CAFEM, mask guidance, and consistency learning under a common data and evaluation protocol.

- iii. Concealment-specific evaluation: performance was reported separately on the official PIDray Easy, Hard, and Hidden test subsets so that improvements on concealed or difficult objects could be distinguished from aggregate detection performance.
- iv. Quantitative explainability and reliability assessment: attention and mask localization were evaluated using intersection-over-union measures, while calibration and Monte Carlo dropout were used to examine the reliability of detector confidence and stochastic uncertainty.
- v. External robustness assessment: the final PIDray-trained model was evaluated zero-shot on an independent SIXray-derived common-class test set, without SIXray fine-tuning or external threshold optimization, to expose the practical effect of cross-dataset domain shift.

The remainder of the manuscript is organized as follows. Section 2 reviews related literature on artificial intelligence for security X-ray inspection, prohibited-item detection, explainable AI, uncertainty and calibration, and cross-dataset domain shift. Section 3 presents the datasets, exploratory data analysis, preprocessing, model architecture, training strategy, explainability and uncertainty methods, and evaluation protocol. Section 4 reports the experimental results, including the B1-B6 ablation, difficulty-stratified PIDray performance, localisation, reliability, and zero-shot SIXray evaluation. Section 5 discusses the findings in relation to prior studies, practical implications, limitations, and future work.

2 Related Literature

2.1 Transformative Impact of Artificial Intelligence on Security X-Ray Screening

Security X-ray inspection is a safety-critical computer-vision task in which prohibited objects must be identified within baggage containing dense, overlapping, and visually heterogeneous contents[19]. Manual inspection remains dependent on operator attention and experience, while automated analysis offers the possibility of faster and more consistent decision support [20],[21]. Deep learning has therefore become increasingly important because it can learn object representations directly from X-ray images rather than relying only on hand-engineered shape or material descriptors[22],[23]. However, the visual characteristics of security X-ray images differ substantially from those of conventional photographic imagery as object boundaries and surface details may be obscured, multiple items may overlap in transmission images, prohibited objects may occupy only a small proportion of the image, and differences in scanner hardware and pseudocoloring schemes can substantially alter the visual appearance and colour distribution of the same material across devices [3],[5].

The SIXray benchmark established several of these challenges at large scale. [10] assembled more than one million security X-ray images and emphasized overlapping objects, complex background structure, extreme positive-negative imbalance, and sparse prohibited-item occurrence as major obstacles to reliable automated inspection. OPIXray later focused directly on occlusion, providing professionally annotated prohibited-item images and separating test cases by occlusion severity

[9]. PIDray extended the field toward deliberately hidden prohibited items by providing 47,677 X-ray images, 12 prohibited-item categories, bounding boxes, segmentation masks, and dedicated Easy, Hard, and Hidden evaluation subsets [8]. Together, these benchmarks shifted prohibited-item detection from relatively simple recognition toward realistic inspection conditions involving concealment, clutter, imbalance, and object overlap.

The evolution of these benchmarks also changes how performance should be interpreted. In a high-security context, a detector that achieves strong average precision but misses a large proportion of hidden items may still be operationally inadequate. Conversely, a highly sensitive model may create an excessive false-alarm burden. Therefore, modern security X-ray research increasingly requires a balance among detection accuracy, difficult-case recall, false-positive control, spatial interpretability, confidence calibration, and robustness to changes in scanners or datasets. This broader view motivates the present study, which evaluates not only mean Average Precision but also Easy-Hard-Hidden performance, localisation consistency, uncertainty, calibration, and zero-shot external transfer.

2.2 Related Literature on Deep Learning for Prohibited-Item X-Ray Detection

Early large-scale deep-learning work in this area concentrated on constructing representative benchmarks and adapting generic visual models to the unusual statistics of X-ray images. Miao et al. [10] introduced SIXray and proposed Class-Balanced Hierarchical Refinement (CHR) to mitigate the dominance of easy negative images and strengthen feature refinement when prohibited items are rare. Their results showed that mid-level representation and class balancing are particularly important when the proportion of positive samples is small. This work established class imbalance and overlapping baggage contents as central research problems rather than incidental dataset noise. Furthermore, [9] subsequently introduced OPIXray and the De-occlusion Attention Module (DOAM). OPIXray was specifically designed around occluded prohibited items and divided test cases by occlusion level. DOAM incorporated edge and material information to guide attention toward partially visible targets. A major conclusion of that study was that the benefit of de-occlusion attention becomes more evident as occlusion increases. This is directly relevant to the present study because concealment-aware modelling should ideally show its largest practical advantage on Hard and Hidden cases rather than only on Easy images.

[8] introduced PIDray as a large-scale benchmark focused on real-world prohibited-item detection, including deliberately hidden objects. Their Selective Dense Attention Network (SDANet) combined dense spatial and channel attention with dependency refinement across multi-scale features. The benchmark demonstrated that attention and multi-scale dependency modelling can improve detection of hidden items, while the accompanying segmentation masks created an opportunity to evaluate not only whether an object is detected but also whether learned evidence is spatially aligned with the prohibited object. PIDray therefore provides the central methodological foundation for the present study.

Subsequent studies have explored different strategies for reducing clutter and recovering occluded targets. [5] proposed foreground-background separation so that irrelevant baggage content could be suppressed before prohibited-item detection. [7] proposed a global context-aware multi-scale feature aggregation method that combines low-level visual cues, spatial attention, global context, and multi-scale fusion for occluded-object detection. These approaches share a common assumption: when a prohibited object is partially hidden, recognition cannot depend only on the strongest local patch; surrounding context and the relationship between object and background become increasingly important.

More recent work has extended X-ray-specific representation learning beyond conventional spatial attention. [4] proposed FDTNet, a dual-stream transformer framework that exploits both spatial/RGB information and frequency-domain information through frequency-aware and global-channel attention. Their findings reinforce the argument that prohibited-item detection benefits from representations tailored to X-ray image characteristics rather than simply transferring generic natural-image features. In parallel, [18] addressed the problem of learning from multiple X-ray datasets through DML-Net, combining appearance enhancement, adversarial prototype alignment, and feature distillation to reduce dataset-specific domain discrepancies while retaining the ability to detect dataset-unique classes.

The literature therefore shows a clear progression from generic deep detection toward problem-specific mechanisms for imbalance, overlap, occlusion, context, and domain discrepancy. However, these contributions are often studied separately. Occlusion-aware attention may improve detection but not quantify whether the resulting explanation is spatially plausible; domain-alignment methods may improve transfer but do not necessarily provide calibrated uncertainty; and explanation methods may be visual without directly affecting concealment-sensitive feature learning. The present study builds on this progression by treating concealment-aware feature refinement, spatial grounding, uncertainty, and external validation as linked components of one inspection framework as shown in table 1.

Table 1 Representative deep-learning studies in prohibited-item X-ray inspection

Study	Dataset / setting	Method or contribution	Key finding / emphasis	Relevance to the present study
[10]	SIXray	Class-Balanced Hierarchical Refinement (CHR)	Established overlap, severe class imbalance, and sparse positive samples as central X-ray inspection challenges.	Motivates evaluation under clutter and cross-dataset transfer.
[9]	OPIXray	De-occlusion Attention	Attention benefits	Supports concealment-

		Module (DOAM)	become more important as prohibited-item occlusion increases.	specific attention and difficulty-stratified testing.
[8]	PIDray	SDANet and large-scale PIDray benchmark	Dense attention and multi-scale dependency refinement improve deliberately hidden-item detection.	Provides the primary dataset, masks, and Easy/Hard/Hidden protocol used here.
[5]	Overlapping X-ray images	Foreground-background separation	Suppressing irrelevant baggage background can improve detection in heavily overlapping scenes.	Relates to CAFEM clutter suppression and object-focused refinement.
[7]	SIXray / OPIXray / WIXray	Global context-aware multi-scale feature aggregation	Combining local cues, attention, global context, and multi-scale fusion improves occluded-object detection.	Supports CAFEM context and multi-scale design rationale.
[4]	OPIXray / PIDray	FDTNet dual-stream transformers	X-ray-specific frequency-aware representations strengthen prohibited-object detection.	Shows value of task-specific feature modelling beyond generic detectors.
[18]	Multi-dataset X-ray detection	DML-Net with prototype alignment and feature distillation	Explicitly targets occlusion and domain discrepancy	Motivates external validation and future domain-

			between security X-ray datasets.	generalisation work.
[17]	PDSXray	Pseudocoloring- driven domain adaptation benchmark	Scanner and pseudocoloring differences can produce substantial cross-domain distribution shift.	Provides a recent basis for interpreting PIDray-to-SIXray transfer limitations.

2.3 Explainable Artificial Intelligence in Prohibited-Item X-Ray Inspection

As automated inspection systems become more accurate, the question of why a detector produces a particular alert becomes increasingly important. In a security setting, an operator may need to know whether a predicted gun, knife, or other prohibited item is supported by a plausible region of the X-ray image rather than by an unrelated background pattern[20]. Post-hoc methods such as Grad-CAM provide heatmaps highlighting which image regions contribute strongly to a neural-network prediction[24],[25]. These visualizations are useful for qualitative inspection, but visual plausibility alone does not establish that the highlighted region corresponds to the actual prohibited object [26].

This limitation is especially important in X-ray images because overlapping objects and pseudo coloring can produce visually salient structures that are unrelated to the prohibited target [5]. A heatmap may therefore appear reasonable while still focusing partly on clutter, edges of benign items, or scanner artefacts [27]. PIDray is valuable in this respect because its segmentation masks and bounding-box annotations create a direct reference against which learned attention or explanation maps can be quantitatively compared[8]. Instead of asking only whether a heatmap looks convincing, localization overlap can be measured using Intersection over Union between an explanation region and an annotated object-support region[28].

The broader explainable-AI literature also argues that explanation should be evaluated rather than presented only as a visual reassurance tool [29]. Grad-CAM established a practical gradient-based approach for model localization[30], while recent application studies in other safety-critical imaging domains have compared explanation maps directly with annotated regions to assess consistency[27]. The methodological implication for security X-ray inspection is that explainability should be connected to observable prohibited-item evidence. The present study therefore evaluates spatial grounding quantitatively and uses mask guidance as part of the model design, rather than treating explanation solely as a post-hoc image overlay.

A further gap is that many prohibited-item detection studies emphasize mAP improvement but give less attention to whether the mechanism producing that improvement is spatially aligned with the target[31],[32]. The present research addresses this limitation by analysing attention IoU and

mask IoU across the progressive B3-B6 variants. This allows explainability to contribute to model evaluation as a measurable property of the detector rather than as an illustrative figure alone.

2.4 Uncertainty Estimation and Reliability-Aware Inspection

Confidence is not equivalent to reliability. Modern neural networks can assign high confidence to incorrect predictions, which is problematic in safety-critical inspection because a missed prohibited item and an unnecessary secondary search have different operational consequences [33]. Monte Carlo Dropout provides a practical Bayesian approximation in which dropout remains active during inference and repeated stochastic forward passes are used to estimate predictive variability [34]. When uncertainty is systematically higher for incorrect than for correct detections, the resulting score can help identify cases that warrant human review.

Calibration is closely related to this problem for example [35] showed that modern deep neural networks can be miscalibrated even when classification accuracy is strong and demonstrated temperature scaling as a simple post-hoc calibration method. In an inspection context, calibration matters because confidence thresholds influence the balance between automated acceptance, alarm generation, and secondary inspection. Expected Calibration Error and the Brier score provide complementary measures of how closely predicted confidence corresponds to empirical correctness[36],[37],[38].

Despite the relevance of these ideas, uncertainty and calibration remain less commonly integrated into prohibited-item X-ray detection than accuracy-oriented architectural improvements. Most studies concentrate on mAP, recall, occlusion handling, or feature representation[39],[40]. This creates a methodological gap where a detector may improve hidden-item recall without providing a reliable indication of when its own predictions are unstable. The present study therefore adds calibration and Monte Carlo dropout analysis to the final detector, treating uncertainty as a separate reliability dimension rather than inferring trustworthiness directly from detection accuracy.

For deployment, this reliability perspective supports a human-in-the-loop workflow rather than fully autonomous inspection. High-confidence, well-localized detections may be handled differently from low-confidence or highly uncertain predictions, while suspicious ambiguous cases can be prioritized for manual review. The present work does not claim that uncertainty alone solves operational decision-making instead, it evaluates whether uncertainty contains meaningful error-discrimination information that could support future referral or triage strategies.

2.5 Domain Shift and Model Generalizability

A detector that performs strongly on a single X-ray dataset may not generalize to images produced by another scanner, pseudo coloring scheme, acquisition environment, annotation convention, or class distribution [17],[41]. This concern is fundamental in security inspection because real deployments use heterogeneous equipment and image-processing pipelines. Cross-dataset evaluation is therefore necessary to distinguish within-domain accuracy from portable model robustness.

[18] explicitly identified large domain discrepancies among security X-ray datasets and showed that naive multi-dataset training can be ineffective when appearance distributions and category sets differ. Their DML-Net addressed this using adversarial prototype alignment for common classes and feature distillation for dataset-specific knowledge. The significance of this work is that domain discrepancy is treated as a first-class modelling problem rather than a minor implementation detail. A detector can learn highly discriminative features within one benchmark while still depending on dataset-specific appearance statistics that fail elsewhere.

[17] extended this concern by introducing PDSXray, a benchmark designed specifically for pseudo coloring-driven and cross-device domain adaptation. Their work demonstrates that security X-ray domain shift can arise not only from object composition and scanner hardware but also from differences in the pseudo coloring process itself. Such shifts can alter color, contrast, material appearance, and background statistics even when the underlying prohibited-item semantics remain similar. These findings provide an important explanation for why external X-ray validation can be substantially more difficult than random splitting within one dataset.

The existing literature therefore suggests two distinct but related robustness problems: concealment within a dataset and distribution shift across datasets. Attention, context aggregation, and mask guidance may strengthen recognition of hidden items inside PIDray, but these mechanisms do not automatically guarantee scanner-independent features. For this reason, the present study uses PIDray for controlled development and then evaluates the final model zero-shot on a SIXray-derived test subset containing only semantically shared classes. This design tests whether the learned representation survives a genuine change in X-ray data distribution without fine-tuning on the external images.

Taken together, the reviewed studies reveal a clear research gap. Existing work has separately advanced prohibited-item detection through class balancing, de-occlusion attention, dense attention, foreground-background separation, global-context aggregation, frequency-aware features, and domain alignment. Explainability and uncertainty methods also exist in the wider deep-learning literature. However, comparatively little work evaluates these dimensions together within a controlled prohibited-item X-ray framework. The present study therefore investigates a concealment-aware, explainable, and uncertainty-aware detector in which (i) feature enhancement is targeted toward difficult and hidden items, (ii) attention is quantitatively compared with object-support regions, (iii) uncertainty and calibration are evaluated explicitly, and (iv) zero-shot external transfer is used to expose the boundary between strong internal detection and cross-domain robustness. This synthesis forms the conceptual basis for the methodology presented in Section 3.

3 Materials and Methods

3.1 Data and Preprocessing

The study used PIDray was used as the primary development and evaluation benchmark because it provides a large collection of security X-ray images together with prohibited-item bounding boxes, segmentation-support annotations, and official Easy, Hard, and Hidden test subsets[8]. The dataset contained 47,677 images in total. The development corpus contained 29,457 images and 39,708 prohibited-item annotations, while the Easy, Hard, and Hidden test subsets jointly contained 18,220 held-out images. Twelve prohibited-item categories were represented: Baton, Pliers, Hammer, Powerbank, Scissors, Wrench, Gun, Bullet, Sprayer, HandCuffs, Knife, and Lighter. Table 2 presents the distribution of the PIDray dataset across the development and held-out test partitions. The development corpus contained 29,457 images with 39,708 prohibited-item annotations, while the Easy, Hard, and Hidden test subsets comprised 9,482, 3,733, and 5,005 images, respectively. All 12 prohibited-item classes were represented across the partitions, while the number of annotations varied considerably with difficulty, particularly in the Hard subset. The table also shows that segmentation information was available for all annotated objects used in the study.

Table 2 PIDray dataset description obtained from the official discovered splits

Split	Images	Annotations	Classes	Segmented annotations
PIDray development corpus	29457	39708	12	39708
PIDray test — Easy	9482	9482	12	9482
PIDray test — Hard	3733	8892	12	8892
PIDray test — Hidden	5005	5008	12	5008

Figure 1 presents representative PIDray X-ray images for the 12 prohibited-item categories used in the study. Each image is displayed with its corresponding bounding-box annotation, class label, and, where available, segmentation support. The visualization demonstrates the diversity of prohibited-item appearance across the dataset and confirms that the data-loading and annotation pipeline correctly preserved the spatial and semantic information required for model development.

These visual checks were performed before training to verify image–annotation correspondence, class-label assignment, bounding-box coordinates, and mask alignment.

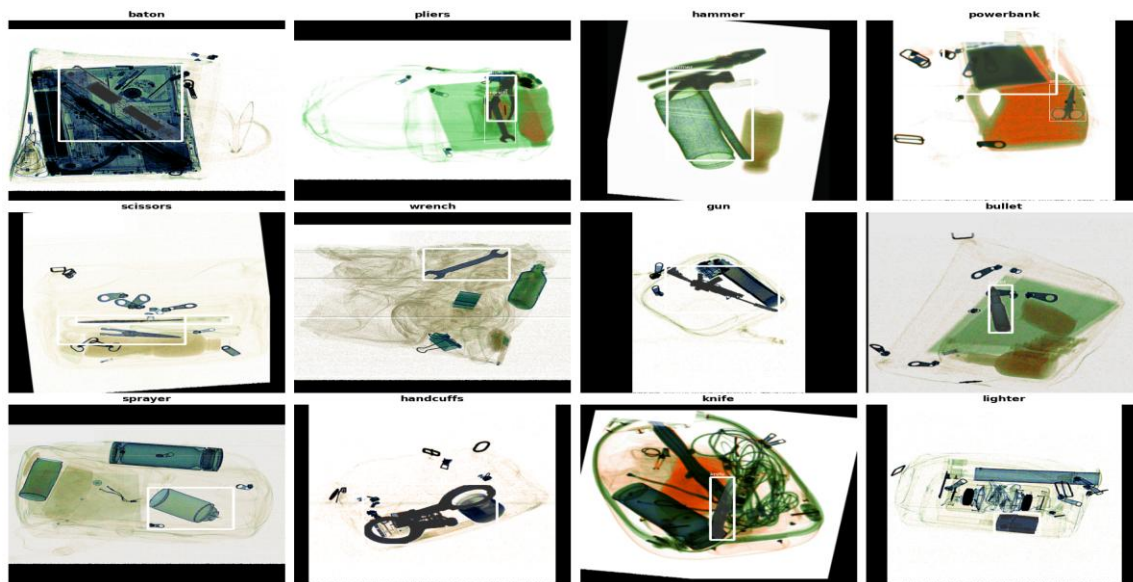


Fig. 1 Representative PIDray X-ray images illustrating the 12 prohibited-item categories with bounding-box annotations and segmentation support.

3.1.1 Exploratory Data Analysis of PIDray

A complete EDA was carried out before model development to establish the statistical and visual properties of the X-ray data. The analysis covered class frequency, class prevalence by image, bounding-box geometry, object density, class co-occurrence, acquisition-related image characteristics, segmentation-area sparsity, and the composition of the official difficulty subsets. The class distribution was not uniform. Powerbank was the most frequent class with 5,171 annotated instances, followed by Scissors (4,352), Wrench (4,350), and Pliers (4,236). Baton was the least frequent class with 1,513 instances. Image prevalence showed the same broad imbalance: Powerbank appeared in approximately 16.98% of development images, while Baton appeared in approximately 4.82%. This variation motivated the use of class-aware stratification for the development split and cautioned against relying on a single aggregate detection metric as shown in table 3 and figures 2 and 3.

Table 3 PIDray prohibited-item class frequency and image prevalence in the development corpus

Class	Instances	Images containing class	Image prevalence (%)
Powerbank	5171	5001	16.98
Scissors	4352	4150	14.09
Wrench	4350	4234	14.37
Pliers	4236	4092	13.89
Lighter	4169	4117	13.98
Hammer	3546	3516	11.94
Knife	3290	3132	10.63
Sprayer	2970	2946	10.00
Gun	2178	2069	7.02
Handcuffs	2096	2091	7.10
Bullet	1837	1832	6.22
Baton	1513	1419	4.82

Bounding -box width (px)	397 08	147.2 608	74.15 80	9.70 87	88.65 25	140.9 384	190.9 091	243.5 780	295.4 128	630.9 677
Bounding -box height (px)	397 08	115.6 333	57.36 24	8.60 93	70.38 66	109.1 510	152.2 936	191.1 887	225.1 223	337.4 065
Bounding -box area / image area	397 08	0.088 9	0.074 6	0.00 05	0.034 9	0.070 9	0.118 3	0.185 0	0.242 4	0.590 1
Aspect ratio (width/height)	397 08	1.654 5	1.323 1	0.06 45	0.773 9	1.284 9	2.120 0	3.255 8	4.209 1	16.85 71

Figures 4 and 5 further illustrate the geometric characteristics of prohibited-item annotations in the PIDray development corpus. Figure 4 shows the distribution of relative bounding-box area, indicating that a substantial proportion of prohibited items occupied only a small fraction of the X-ray image. The median bounding box covered approximately 7.09% of the image area, while the lower quartile covered approximately 3.49%, confirming the presence of many relatively small targets. Figure 5 presents the distribution of bounding-box aspect ratios and demonstrates considerable variation in object shape. The median width-to-height ratio was approximately 1.28, while the 95th percentile reached approximately 4.21 and the maximum exceeded 16.8. These distributions highlight the considerable variation in both object size and shape within PIDray and support the use of multi-scale detection features and context-aware feature refinement.

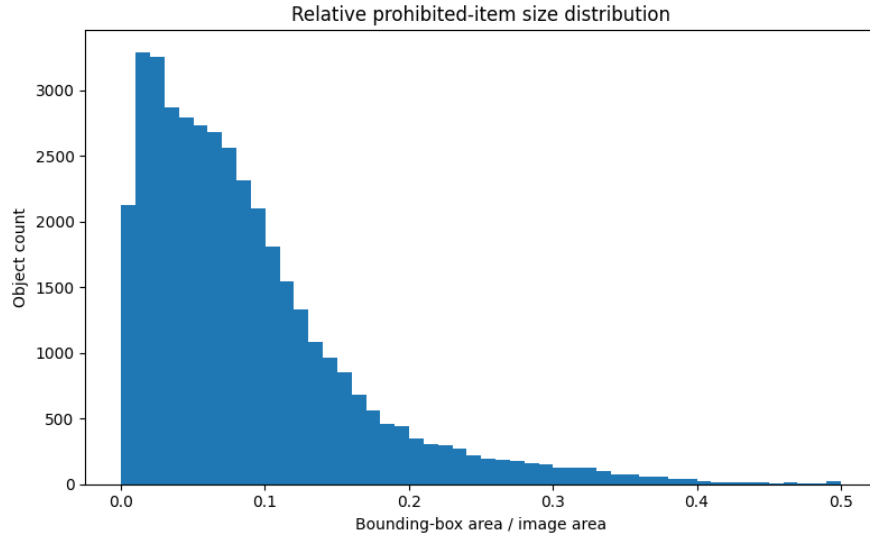


Fig. 4 Distribution of prohibited-item bounding-box area relative to image area.

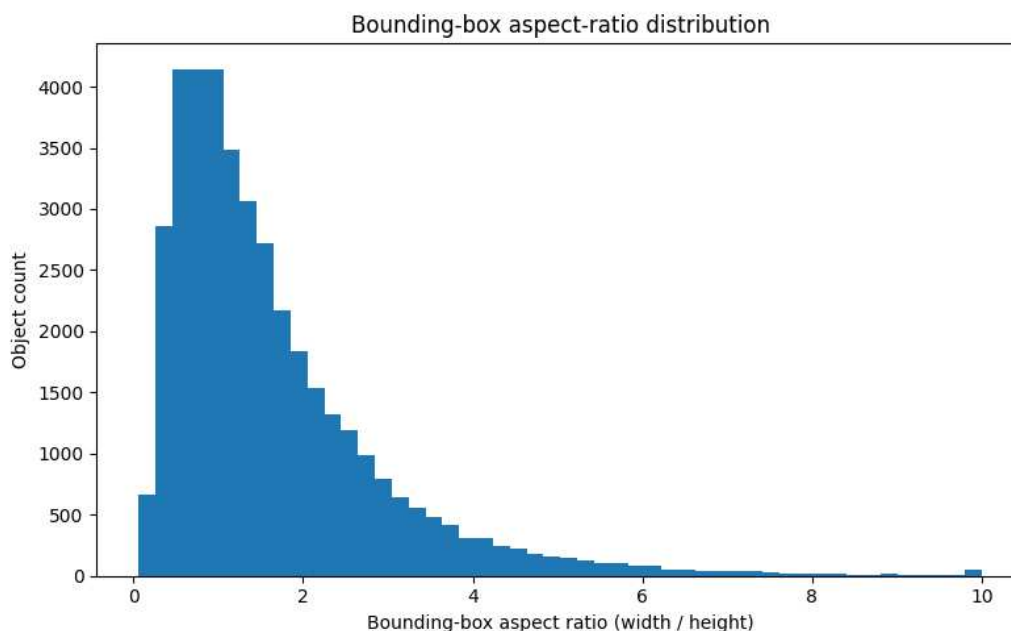


Fig. 5 PIDray prohibited-item bounding-box aspect-ratio distribution

Object density averaged approximately 1.35 annotated prohibited items per development image, with a median of one object and a 75th percentile of two objects. Although most images therefore contained one or two prohibited items, the maximum reached 25, indicating that a subset of X-rays contained substantially more complex object arrangements as shown in table 5 and figure 6. This motivates evaluation under clutter and overlap rather than treating every image as a single-object scene.

Table 5 Descriptive statistics for prohibited-item object density per development image

Statistic	Objects per image
count	29457
mean	1.347999
std	0.697576
min	0
25%	1
50%	1
75%	2
max	25

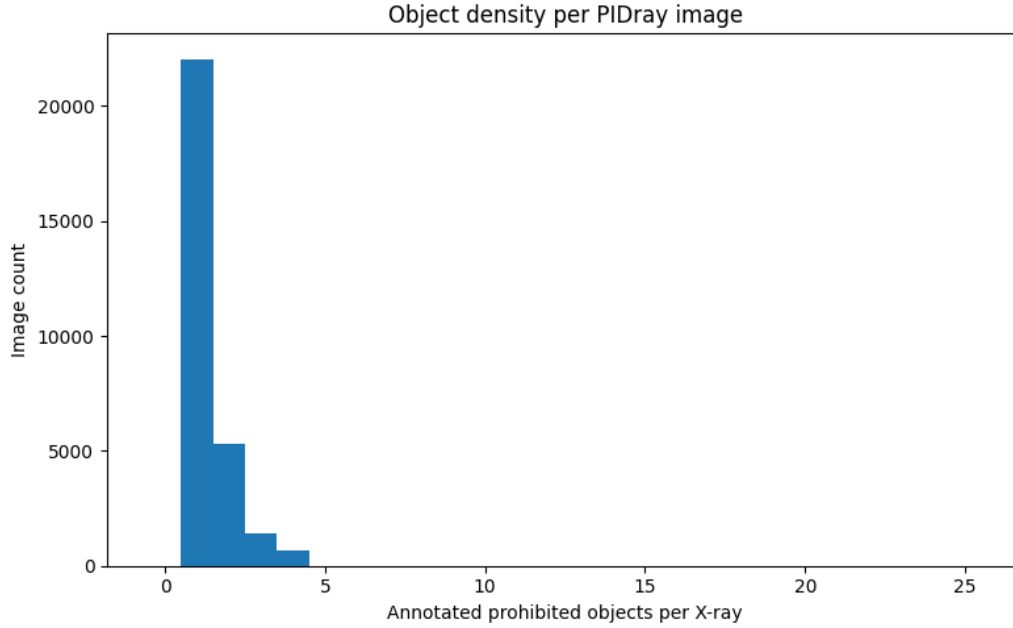


Fig. 6 Distribution of annotated prohibited-object density per PIDray development image.

The co-occurrence matrix shows that some prohibited-item classes frequently appeared together. Among the larger off-diagonal counts were Pliers-Wrench (1,209 images), Scissors-Knife (1,185), Pliers-Hammer (1,110), Powerbank-Scissors (1,112), and Sprayer-Lighter (908). Such co-occurrence is relevant to concealment-aware detection because overlapping or neighbouring objects can obscure edges and alter the local appearance of a prohibited item as shown in table 6.

Table 6 PIDray class co-occurrence matrix by image

Class	Bat	Pli	Ha m	Pow	Sci	Wre	Gun	Bul	Spr	Han	Kni	Lig
Baton	141 9	0	0	0	0	1	96	98	0	426	0	134
Pliers	0	409 2	1110	374	123	120 9	0	0	0	0	118	0
Hammer	0	111 0	3516	147	124	987	0	0	0	0	119	0
Powerban k	0	374	147	500 1	111 2	373	0	0	655	0	431	510
Scissors	0	123	124	111 2	415 0	0	0	0	295	0	118 5	147
Wrench	1	120 9	987	373	0	423 4	0	0	0	1	0	0
Gun	96	0	0	0	0	0	206 9	548	0	484	0	0
Bullet	98	0	0	0	0	0	548	183 2	0	489	0	0

Sprayer	0	0	0	655	295	0	0	0	294 6	0	148	908
Handcuffs	426	0	0	0	0	1	484	489	0	209 1	0	134
Knife	0	118	119	431	118 5	0	0	0	148	0	313 2	0
Lighter	134	0	0	510	147	0	0	0	908	134	0	411 7

Column abbreviations: Bat=Baton; Pli=Pliers; Ham=Hammer; Pow=Powerbank; Sci=Scissors; Wre=Wrench; Bul=Bullet; Spr=Sprayer; Han=HandCuffs; Kni=Knife; Lig=Lighter.

Image-quality EDA was performed on a reproducible random sample of 1,000 development images. Image dimensions were heterogeneous: the median width was 420 pixels and the median height was 448 pixels, with widths extending to 2,000 pixels and heights to 1,040 pixels. Mean grayscale brightness had a median of approximately 226.81, while grayscale contrast had a median standard deviation of approximately 55.70. The Laplacian-variance blur proxy also varied widely, indicating that the dataset contains substantial acquisition-level diversity. These observations support standardized resizing and consistent evaluation preprocessing. Table 7 summarizes the descriptive statistics for image dimensions, brightness, contrast, and sharpness-related measures used in the quality assessment. Figures 7–11 provide the corresponding visual distributions of these image-quality characteristics, showing variation in image width, height, brightness, contrast, and the Laplacian-variance blur proxy across the sampled images. The distributions demonstrate substantial heterogeneity in acquisition and image quality, supporting the use of standardized resizing and consistent preprocessing before model development and evaluation.

Table 7 Image-quality and acquisition statistics from the 1,000-image PIDray EDA sample

Measure	n	Mean	SD	Min	25%	Median	75%	Max
Image width (px)	100 0	477.588	221.855	180.00 0	360.000	420.000	540.000	2000.00 0
Image height (px)	100 0	481.744	137.320	448.00 0	448.000	448.000	448.000	1040.00 0
Mean grayscale brightness	100 0	223.633	14.755	164.75 8	218.453	226.811	233.578	249.016
Grayscale contrast SD	100 0	56.779	14.524	22.202	46.402	55.697	66.971	102.035
Laplacian variance	100 0	3041.69 0	1869.94 7	141.40 6	1120.08 5	3272.22 5	4368.23 3	9110.18 4

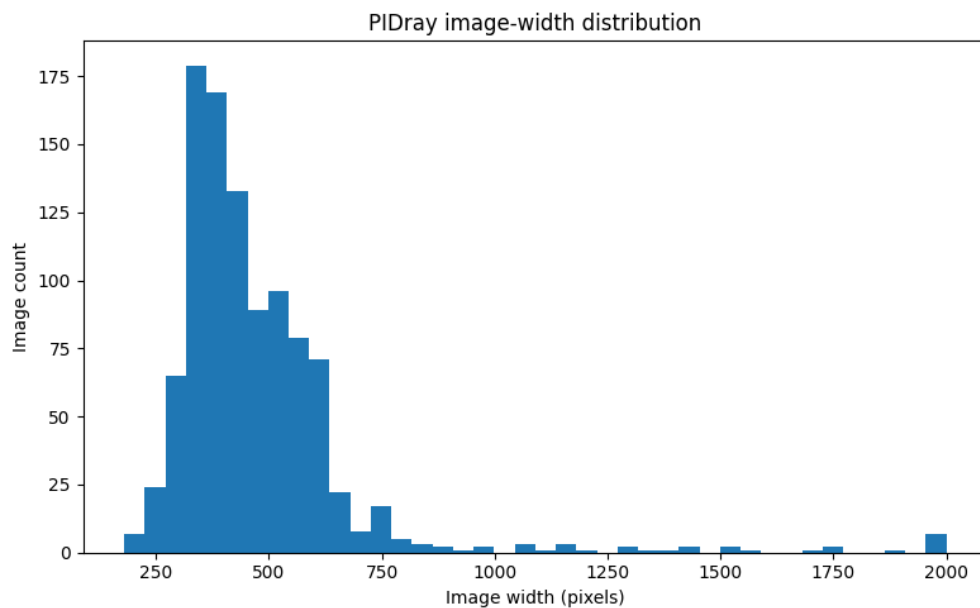


Fig. 7 PIDray image-width distribution in the 1,000-image EDA sample

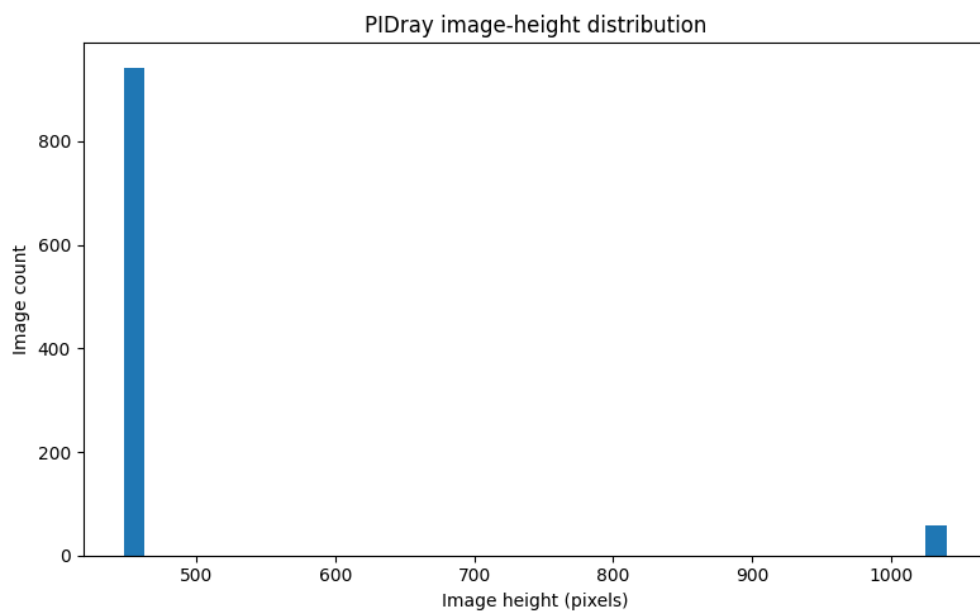


Fig. 8 PIDray image-height distribution in the 1,000-image EDA sample

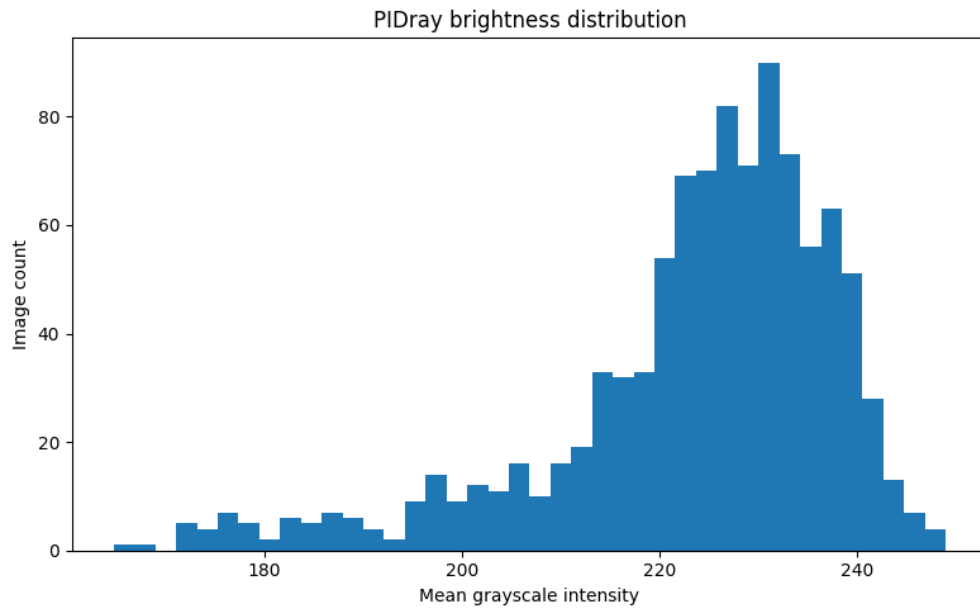


Fig. 9 Distribution of mean grayscale brightness in the PIDray EDA sample.

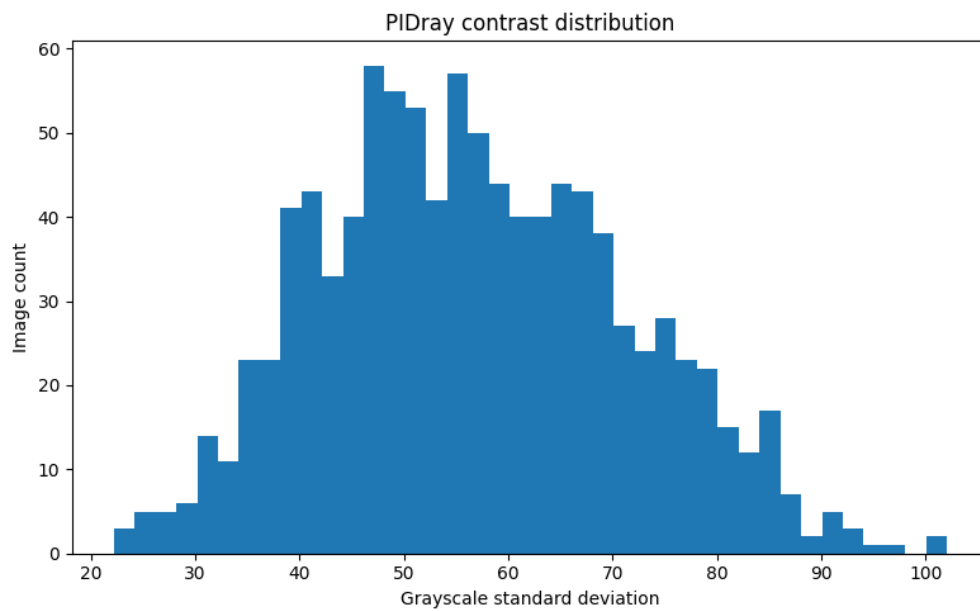


Fig. 10 Distribution of grayscale contrast in the PIDray EDA sample.

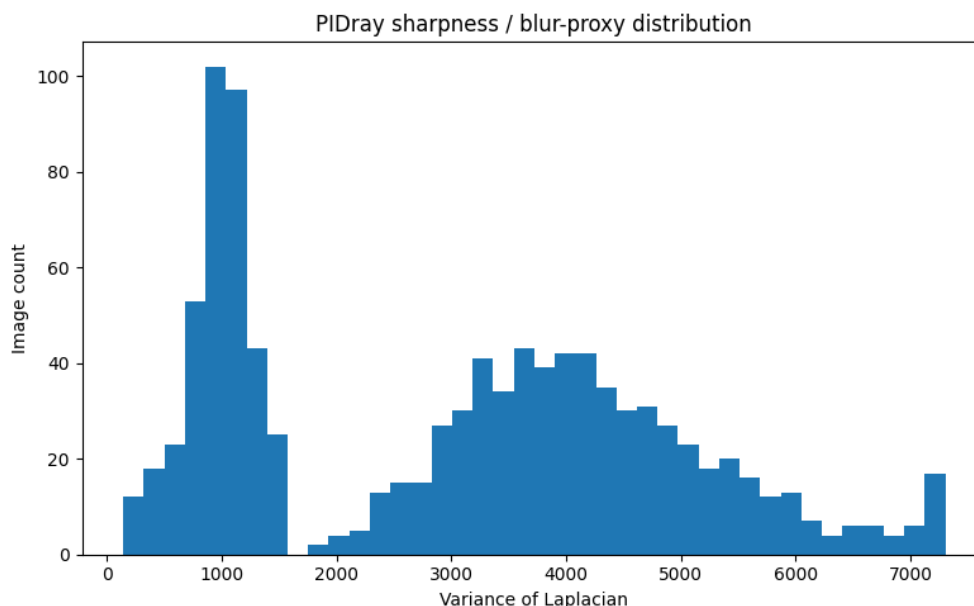


Fig. 11 PIDray sharpness/blur-proxy distribution based on variance of the Laplacian.

To characterize the sparsity of the segmentation-support annotations, the analysis computed an annotation-area fraction using the COCO annotation area where available. Table 8 summarizes the descriptive statistics of this annotation-area fraction, which was used as a proxy for segmentation sparsity within the development corpus. The resulting values were concentrated toward very small fractions, indicating that the annotated pixel-support regions generally occupied only a limited proportion of each X-ray image.

Table 8 Summary of the annotation-area fraction used as a segmentation sparsity proxy

Statistic	Annotation area / image area
count	39,708
mean	5.071359e-06
std	1.614335e-06
min	4.807692e-07
25%	3.985969e-06
50%	5.073052e-06
75%	6.200397e-06
90%	6.975446e-06
95%	7.440476e-06
99%	8.585165e-06
max	1.594388e-05

Figure 12 visualizes the distribution of the annotation-area fraction and further illustrates the strong concentration of observations at low values. Because this measure depends on the area metadata encoded in the source annotations, it was interpreted descriptively rather than as a direct physical estimate of the visible proportion of a prohibited item. Its main purpose was to document the

sparsity of the available pixel-level support used by the mask-guided components of the proposed framework.

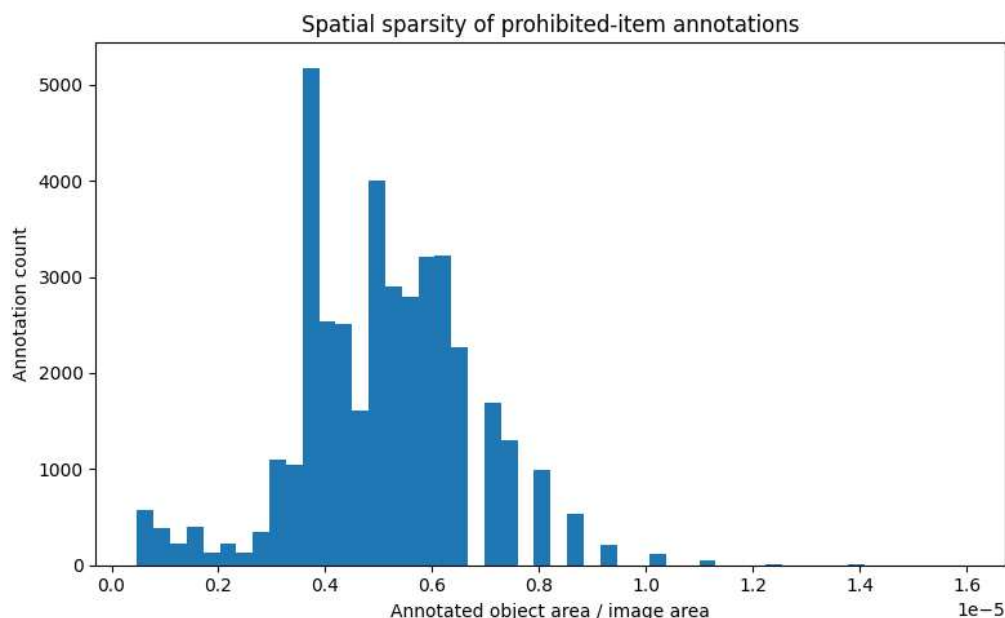


Fig. 12 Spatial sparsity distribution of prohibited-item annotation area in the PIDray development corpus.

To examine the structural differences among the official PIDray test subsets, the number of images, annotations, and average objects per image were compared across Easy, Hard, and Hidden conditions. Table 9 summarizes the composition of the three official test subsets and highlights differences in scene complexity and object density. Easy contained 9,482 images with approximately one annotation per image, while Hidden contained 5,005 images with a similar average object count. In contrast, Hard contained 3,733 images and 8,892 annotations, corresponding to approximately 2.38 objects per image.

Table 9 Composition of the official PIDray Easy, Hard, and Hidden test subsets.

Difficulty	Images	Annotations	Objects per image
Easy	9482	9482	1.0000
Hard	3733	8892	2.3820
Hidden	5005	5008	1.0006

Figure 13 provides a visual comparison of the Easy, Hard, and Hidden test-set composition, further illustrating the substantially higher annotation density in the Hard subset. This pattern indicates that test difficulty is associated not only with object appearance or concealment but also with more

complex multi-object scenes. For this reason, performance on the Easy, Hard, and Hidden subsets was evaluated and reported separately in Section 4.

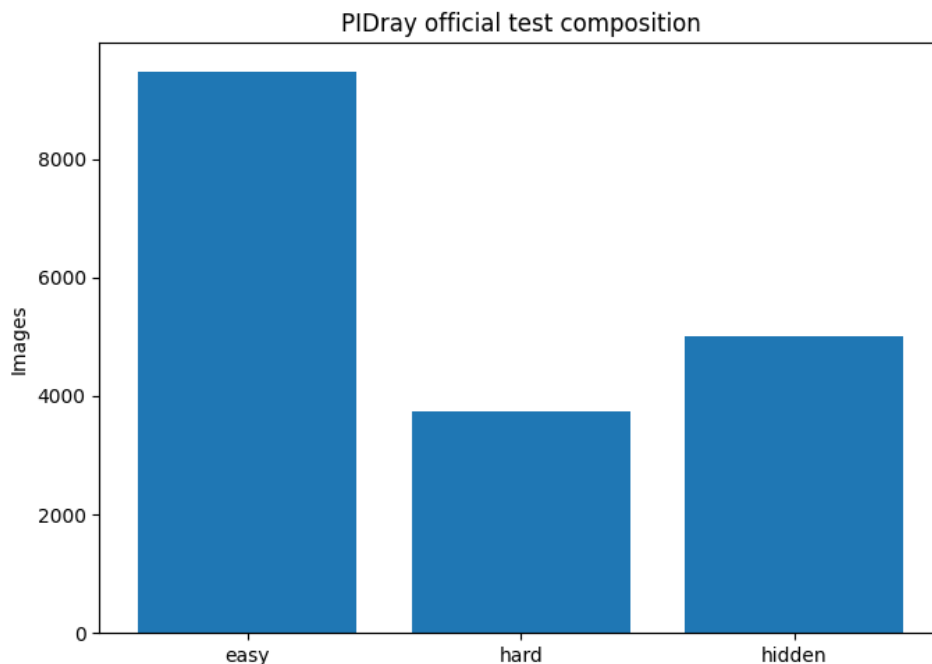


Fig. 13 Number of images in the official PIDray Easy, Hard, and Hidden test subsets

3.1.2 Development Split, Preprocessing, and Data-Leakage Control

The 29,457-image PIDray development corpus was partitioned using iterative multi-label stratification into 26,510 training images (90%) and 2,947 validation images (10%). Each image was represented by the set of prohibited-item classes it contained as shown in table 10. The purpose of multi-label stratification was to preserve class prevalence in both partitions while preventing overlap between training and validation. The split produced zero train-validation image overlap.

Table 10 Class prevalence after the multi-label stratified 90:10 development split

Class	Training images	Validation images	Train prevalence (%)	Validation prevalence (%)	Absolute difference (pp)
Baton	1277	142	4.817	4.819	0.001
Pliers	3683	409	13.893	13.878	0.014
Hammer	3164	352	11.935	11.944	0.009
Powerbank	4501	500	16.979	16.966	0.012
Scissors	3735	415	14.089	14.082	0.007
Wrench	3811	423	14.376	14.354	0.022
Gun	1862	207	7.024	7.024	0.000
Bullet	1649	183	6.220	6.210	0.011
Sprayer	2651	295	10.000	10.010	0.010
Handcuffs	1882	209	7.099	7.092	0.007
Knife	2819	313	10.634	10.621	0.013

Lighter	3705	412	13.976	13.980	0.004
---------	------	-----	--------	--------	-------

The maximum absolute difference in class prevalence between the training and validation partitions was only 0.00022 in proportion units (approximately 0.022 percentage points), demonstrating that the stratified split preserved class representation closely. The official Easy, Hard, and Hidden test sets remained untouched throughout model development.

All images were standardized to the input resolution used by the final controlled CAFEM experiments (640×640 pixels). Image and annotation transformations were applied jointly to maintain geometric consistency between the X-ray, bounding boxes, and masks. Training augmentation was deliberately conservative because security X-ray images encode material density, overlap, and low-contrast structure differently from ordinary photographic imagery. Validation and test preprocessing remained deterministic. Figure 14 provides the generated visual sanity check of evaluation preprocessing and representative training augmentations.

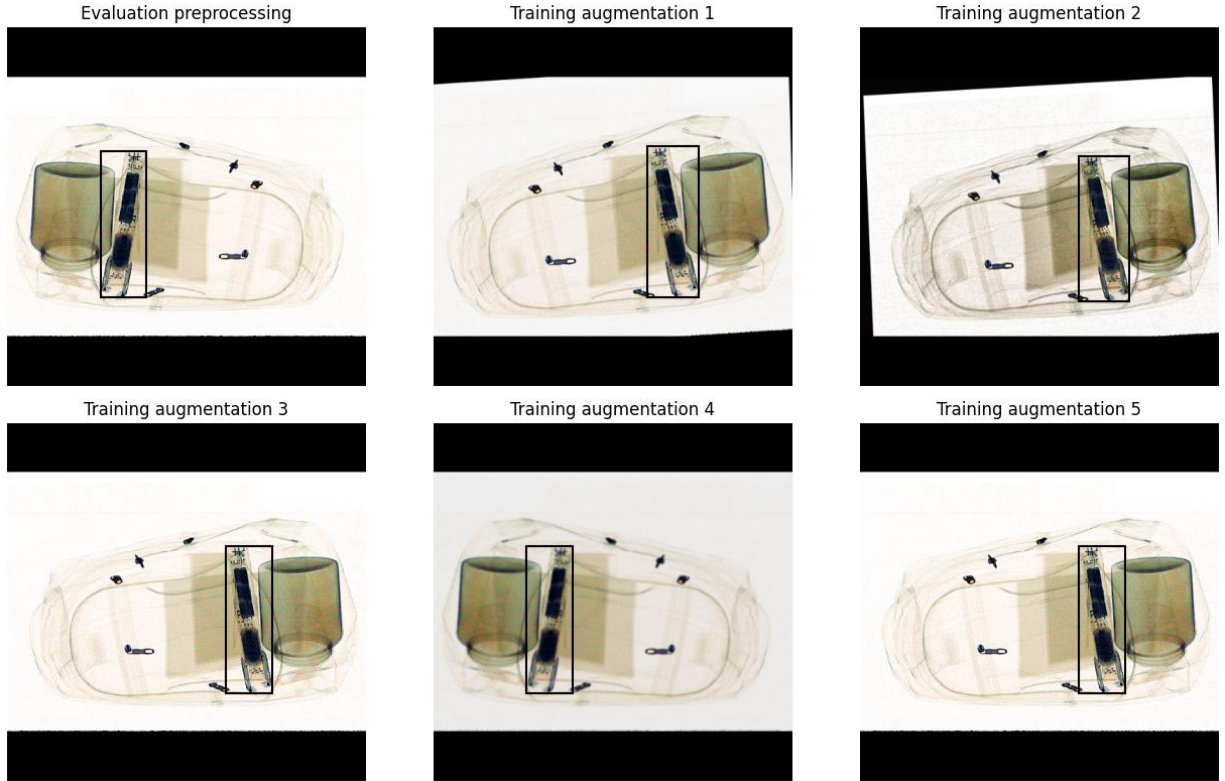


Fig. 14 Evaluation preprocessing and representative training augmentations visualized on a PIDray image

3.1.3 SIXray-Derived External Dataset

External robustness was evaluated using an independent SIXray-derived test subset [10] to determine how well the PIDray-trained detector transferred to a different security X-ray data distribution without additional fine-tuning. The external subset comprised 831 X-ray images with

831 corresponding annotation files and was restricted to five prohibited-item categories that were semantically shared with PIDray: Gun, Knife, Pliers, Scissors, and Wrench. Table 11 presents the class mapping used to align the SIXray labels with the corresponding PIDray class labels for zero-shot external evaluation. Restricting the evaluation to these common classes ensured that the external test measured cross-dataset generalization rather than differences caused by incompatible label spaces.

Table 11 Common-class mapping used for the SIXray-derived zero-shot evaluation

SIXray class ID	External class	PIDray label	PIDray class
0	Gun	7	Gun
1	Knife	11	Knife
2	Pliers	2	Pliers
3	Scissors	5	Scissors
4	Wrench	6	Wrench

3.2 Concealment-Aware Feature Enhancement Module

The principal methodological component of CAU-XInspect is the Concealment-Aware Feature Enhancement Module (CAFEM). CAFEM was inserted immediately before the YOLO26 Detect head and operated independently on each incoming detection scale. For a detection feature map F_s at scale s , a learnable scale-specific projection operator, P_s , was first applied to transform the original feature channels into a common hidden representation. In this study, P_s denotes a 1×1 convolution that adjusts the channel dimension while preserving the spatial resolution of the feature map. The projected representation Z_s was therefore defined as:

$$Z_s = P_s(F_s) \quad (1)$$

where $F_s \in \mathbb{R}^{C_s \times H_s \times W_s}$ is the input feature map at detection scale s , P_s is the learnable 1×1 convolutional projection for that scale, and $Z_s \in \mathbb{R}^{C_h \times H_s \times W_s}$ is the resulting hidden feature representation with C_h projected channels.

Three complementary feature-processing branches were subsequently applied to the projected representation Z_s . These branches were designed to capture different forms of evidence that may remain informative when a prohibited item is partially occluded, weakly contrasted, or embedded within surrounding baggage clutter.

The occlusion branch applied a 3×3 depthwise convolution followed by a pointwise convolution to model local spatial structures while retaining computational efficiency. Its output was defined as

$$O_s = G_s^{\text{occ}}(Z_s) = \text{PWConv}_s^{\text{occ}}\left(\text{DWConv}_{3 \times 3, s}^{\text{occ}}(Z_s)\right), \quad (2)$$

where O_s denotes the occlusion-sensitive feature representation at scale s ; G_s^{occ} denotes the complete occlusion-processing branch; $\text{DWConv}_{3 \times 3, s}^{\text{occ}}$ is a 3×3 depthwise convolution that extracts local spatial patterns independently within each channel and $\text{PWConv}_s^{\text{occ}}$ is a 1×1 pointwise convolution used to combine the resulting channel-wise information. This branch was intended to preserve local structural cues that may remain visible even when only part of a prohibited item is exposed.

The contrast branch first constructed a local residual by subtracting a 3×3 average-pooled representation from the projected feature map. The residual was then processed by the branch-specific convolutional operator:

$$\begin{aligned} R_s^{\text{ctr}} &= Z_s - \text{AvgPool}_{3 \times 3}(Z_s), \\ C_s &= G_s^{\text{ctr}}(R_s^{\text{ctr}}), \end{aligned} \quad (3)$$

where R_s^{ctr} represents the local contrast residual, $\text{AvgPool}_{3 \times 3}(\cdot)$ denotes 3×3 average pooling, G_s^{ctr} denotes the learnable contrast-processing operator, and C_s is the resulting contrast-sensitive representation. By suppressing locally averaged responses, this branch emphasizes deviations from the surrounding region and therefore helps preserve weak object boundaries and subtle local intensity differences.

The context branch was designed to capture a broader receptive field using a dilated depth wise convolution followed by pointwise channel mixing:

$$X_s = G_s^{\text{ctx}}(Z_s) = \text{PWConv}_s^{\text{ctx}} \left(\text{DWConv}_{3 \times 3, d, s}^{\text{ctx}}(Z_s) \right), \quad (4)$$

where X_s denotes the contextual feature representation; G_s^{ctx} is the contextual-processing branch; $\text{DWConv}_{3 \times 3, d, s}^{\text{ctx}}$ denotes a 3×3 depthwise convolution with dilation rate d ; and $\text{PWConv}_s^{\text{ctx}}$ denotes the corresponding 1×1 pointwise convolution. The dilation enlarges the effective receptive field without reducing the spatial resolution of the feature map, enabling the branch to incorporate surrounding evidence that may assist recognition when the prohibited item itself is only partially visible.

The outputs of the three complementary branches were then concatenated along the channel dimension and transformed by a learnable fusion operator:

$$H_s = G_s^{\text{fuse}}([O_s \parallel C_s \parallel X_s]), \quad (5)$$

where $[\cdot \parallel \cdot \parallel \cdot]$ denotes channel-wise concatenation, G_s^{fuse} denotes the learnable fusion operator, and H_s represents the fused concealment-aware feature representation at scale s . This operation

allows local occlusion cues, contrast information, and broader contextual evidence to be integrated before attention weighting.

A spatial attention map was subsequently generated from the fused representation:

$$A_s = \sigma(G_s^{\text{att}}(H_s)), \quad (6)$$

where G_s^{att} denotes the learnable attention-projection operator, $A_s \in [0, 1]^{1 \times H_s \times W_s}$ is the single-channel spatial attention map, and $\sigma(\cdot)$ is the sigmoid activation function that constrains the attention weights to the interval $[0, 1]$. Higher values of A_s indicate spatial locations receiving greater emphasis during concealment-aware feature refinement.

Finally, the original YOLO26 detection feature map was refined using a gated residual update:

$$\tilde{F}_s = F_s + \gamma_s [\Delta_s(H_s) \odot A_s], \quad (7)$$

where $\tilde{F}_s$ denotes the refined feature map passed to the YOLO26 Detect head, $\Delta_s(\cdot)$ is a learnable projection that transforms the fused representation to the channel dimensionality required by the original detection feature map, A_s is the spatial attention map from Equation 6, $\odot$, denotes element-wise multiplication with spatial broadcasting across channels, and γ_s is a learnable residual-scaling coefficient controlling the contribution of the CAFEM refinement.

3.2.1 Mask Guidance and Consistency Learning

For the mask-guided variants, CAFEM incorporated an auxiliary mask-prediction head to provide explicit spatial guidance during feature refinement. At each detection scale s , the fused concealment-aware representation H_s was passed through a learnable mask-prediction operator and a sigmoid activation to produce a probability map:

$$M_s = \sigma(G_s^{\text{mask}}(H_s)), \quad (8)$$

where G_s^{mask} denotes the learnable mask-prediction head at scale s , H_s is the fused concealment-aware feature representation obtained from Equation 5, $\sigma(\cdot)$ is the sigmoid activation function, and $M_s \in [0, 1]^{1 \times H_s \times W_s}$ denotes the predicted mask-probability map. Higher values of M_s indicate spatial locations predicted to correspond more strongly to prohibited-item regions.

The predicted mask was then used to multiplicatively gate the refined detection features:

$$\hat{F}_s = \tilde{F}_s \odot (1 + \gamma_m M_s), \quad (9)$$

where $\tilde{F}_s$ is the CAFEM-refined feature map obtained from Equation 7, $\hat{F}_s$ is the mask-guided feature map passed toward the detection head, γ_m is a learnable mask-scaling parameter controlling the strength of mask guidance, and $\odot$, denotes element-wise multiplication with the single-channel mask map broadcast across the feature channels. The additive term $1 + \gamma_m M_s$ preserves the original feature response while increasing the contribution of spatial locations emphasized by the auxiliary mask.

Supervision of the predicted attention and mask maps used a common map-supervision objective combining binary cross-entropy (BCE) and soft Dice loss. Let $L = \{L_i\}_{i=1}^N$ denote the predicted logits for a spatial map and $Y = \{Y_i\}_{i=1}^N$ the corresponding binary target map, where N is the number of spatial locations. The predicted probability at location i was obtained as

$$p_i = \sigma(L_i).$$

The binary cross-entropy component was defined as

$$\mathcal{L}_{\text{BCE}}(L, Y) = -\frac{1}{N} \sum_{i=1}^N [Y_i \log(p_i) + (1 - Y_i) \log(1 - p_i)] \quad (10)$$

where $p_i = \sigma(L_i)$ is the predicted probability at spatial location i , $Y_i \in \{0, 1\}$ is the corresponding target value, and N denotes the total number of pixels or spatial positions included in the supervision map. This component penalizes incorrect pixel-level foreground and background predictions.

To complement BCE and account for the sparsity of prohibited-item support regions, a soft Dice loss was also computed:

$$\mathcal{L}_{\text{Dice}}(L, Y) = 1 - \frac{2 \sum_{i=1}^N p_i Y_i + \varepsilon}{\sum_{i=1}^N p_i + \sum_{i=1}^N Y_i + \varepsilon} \quad (11)$$

where p_i and Y_i retain the definitions given above and ε is a small numerical-stability constant that prevents division by zero. The Dice component emphasizes spatial overlap between the predicted map and the target region and is particularly useful when positive pixels occupy only a small proportion of the image.

The complete map-supervision loss was therefore defined as

$$\mathcal{L}_{\text{map}} = 0.5 \mathcal{L}_{\text{BCE}} + 0.5 \mathcal{L}_{\text{Dice}} \quad (12)$$

where $\mathcal{L}_{\text{map}}$ represents the combined supervision applied to the relevant attention or mask prediction, $\mathcal{L}_{\text{BCE}}$ provides pixel-wise classification supervision, and $\mathcal{L}_{\text{Dice}}$ encourages spatial overlap with the annotated prohibited-item support region. Using the two terms together reduced the risk that the learning objective would be dominated by the large background area present in sparse X-ray annotations.

For the full B6 configuration, an additional consistency objective was introduced to encourage agreement between the independently generated attention and mask probability maps. Consistency was implemented using a Smooth-L1 penalty rather than an IoU-based loss, in accordance with the final experimental implementation:

$$\mathcal{L}_{\text{cons}} = \frac{1}{N} \sum_{i=1}^N \text{SmoothL1}(A_{s,i}, \tilde{M}_{s,i}) \quad (13)$$

where $A_{s,i}$ denotes the attention probability at spatial location i from Equation 6, $\tilde{M}_{s,i}$ denotes the corresponding mask probability after resizing M_s to the spatial dimensions of the attention map when necessary, N is the number of compared spatial locations, and $\text{SmoothL1}(\cdot)$ denotes the Smooth-L1 penalty. The consistency term encourages the attention and mask mechanisms to emphasize similar prohibited-item regions while remaining less sensitive to occasional large deviations than a purely squared-error objective.

3.2.2 Training-Only Concealment Proxy

The official PIDray Easy, Hard, and Hidden labels were reserved exclusively for held-out evaluation and were not used during model training. Instead, the final CAFEM experiment derived a training-only concealment proxy from three measurable properties of each annotated object: relative object size, mask fill, and same-image crowding. This design allowed the model to emphasize potentially difficult objects without introducing information from the official test-difficulty labels into the training process.

For object i , the relative bounding-box area was first defined as the bounding-box area divided by the total image area and denoted by r_i . A smallness score was then computed as

$$s_i = 1 - \min\left(1, \sqrt{\frac{r_i}{0.04}}\right), \quad (14)$$

where $s_i \in [0, 1]$ denotes the smallness score for object i , r_i is the fraction of the image occupied by its bounding box, and 0.04 represents the reference area corresponding to approximately 4% of the image. Objects occupying 4% or more of the image therefore receive little or no smallness emphasis, whereas progressively smaller objects receive larger s_i values. The square-root transformation moderates the rate at which the score increases as object area decreases.

A complementary low-fill score was computed from the proportion of the bounding box occupied by the corresponding segmentation mask:

$$f_i = 1 - \min\left(1, \sqrt{\frac{\text{maskArea}_i}{\text{boxArea}_i}}\right), \quad (15)$$

where $f_i \in [0, 1]$ denotes the low-fill score, maskArea_i is the number of pixels belonging to the object's segmentation support, and boxArea_i is the area of its bounding box. A low mask-to-box area ratio produces a larger f_i , indicating that only a small fraction of the bounding region contains object-support pixels. This characteristic was used as a proxy for weakly filled or spatially sparse object evidence.

The third component measured same-image crowding by determining the largest bounding-box overlap between object i and any other annotated object in the same image:

$$q_i = \max_{j \neq i} \text{IoU}(b_i, b_j), \quad (16)$$

where $q_i \in [0, 1]$ denotes the crowding score for object i , b_i is its bounding box, b_j is the bounding box of another object j in the same image, and $\text{IoU}(b_i, b_j)$ is their Intersection over Union. A larger q_i indicates greater overlap with neighbouring objects and therefore a more spatially cluttered detection context.

The three object-level characteristics were then combined to form the training-only concealment proxy:

$$c_i = 0.40s_i + 0.35f_i + 0.25q_i, \quad (17)$$

where $c_i \in [0, 1]$ denotes the composite concealment-proxy score for object i . The coefficients assign the greatest contribution to object smallness, followed by low mask fill and then crowding. Consequently, objects that are simultaneously small, weakly filled, and strongly overlapped receive larger proxy scores and are treated as potentially more difficult during training.

The object-level proxy was subsequently transformed into a difficult-object target intensity:

$$t_i = 0.50 + 0.50c_i, \quad (18)$$

where $t_i \in [0.5, 1.0]$ denotes the target intensity associated with object i . This formulation maintains a minimum supervision intensity of 0.5 while progressively increasing the target contribution as the concealment-proxy score approaches 1.0.

Finally, the object-level proxy scores were aggregated to compute a batch-level detector weighting factor:

$$w_{\text{det}} = 1 + 0.25 \left(\frac{1}{N} \sum_{i=1}^N c_i \right), \quad (19)$$

where w_{det} denotes the detector-loss weighting factor, N is the number of annotated object instances contributing to the current batch-level calculation, and $\frac{1}{N} \sum_{i=1}^N c_i$ is the mean concealment-proxy score across those objects. As the average training difficulty increases, w_{det} increases moderately above 1.0, thereby giving greater emphasis to the detector loss for batches containing more difficult object configurations.

3.2.3 Composite Training Objective

The detector loss remained the principal optimization objective, while auxiliary losses were introduced progressively according to the controlled ablation design. For the complete B6 configuration, the total training objective was defined as

$$\mathcal{L}_{\text{total}} = w_{\text{det}} \mathcal{L}_{\text{det}} + 0.20 \mathcal{L}_{\text{conceal}} + 0.30 \mathcal{L}_{\text{seg}} + 0.10 \mathcal{L}_{\text{cons}}, \quad (20)$$

where $\mathcal{L}_{\text{total}}$ denotes the overall optimization loss, $\mathcal{L}_{\text{det}}$ is the standard YOLO26 detection loss, and w_{det} is the training-only difficulty-dependent weighting factor defined in Equation 19. The term $\mathcal{L}_{\text{conceal}}$ denotes the concealment-aware map-supervision loss used to encourage CAFEM to emphasize spatial regions associated with potentially difficult prohibited items, while $\mathcal{L}_{\text{seg}}$ represents the auxiliary segmentation loss used to supervise the predicted mask against the available prohibited-item mask support. The term $\mathcal{L}_{\text{cons}}$ is the Smooth-L1 consistency loss defined in Equation 13, which encourages agreement between the attention and mask probability maps.

The coefficients 0.20, 0.30, and 0.10 control the relative contributions of concealment supervision, segmentation supervision, and attention–mask consistency, respectively. The detector loss retained the dominant role in optimization, while the auxiliary objectives acted as regularizing spatial constraints rather than replacing the principal object-detection objective.

The auxiliary terms were activated progressively across the ablation variants. B3 used a supervised attention loss weighted by 0.15 in place of the CAFEM concealment term. B4 introduced the CAFEM concealment-map loss, B5 additionally incorporated segmentation supervision, and B6 further added the Smooth-L1 consistency term. Thus, each experimental variant introduced only the auxiliary objective associated with the corresponding architectural component, enabling the contribution of attention, concealment-aware refinement, mask guidance, and consistency learning to be evaluated under a controlled training protocol.

3.3 Concealment-Aware Prohibited-Item Detection

3.3.1 Controlled Baseline and Ablation Design

YOLO26 served as the controlled B1 reference. Enhanced variants were initialized from the same original YOLO26 checkpoint. The backbone and neck were frozen, while the inserted module and final Detect head were fine-tuned where required. This design minimized confounding from complete retraining of the base detector and allowed the contribution of each added component to be studied under a common detector, data split, and evaluation protocol as shown in Table 12.

Table 12 Controlled B1-B6 ablation configurations

Variant	Configuration	Purpose
B1	YOLO26	Controlled detector baseline
B2	YOLO26 + CBAM	Conventional attention comparator
B3	YOLO26 + CAAM	Supervised concealment-aware attention comparator
B4	YOLO26 + CAFEM	Proposed multi-branch concealment feature enhancement
B5	YOLO26 + CAFEM + mask guidance	Tests contribution of object-support mask guidance
B6	YOLO26 + CAFEM + mask guidance + consistency	Full concealment-aware model and reliability-analysis model

3.4 Explainability and Uncertainty

Explainability and uncertainty were treated as measurable reliability properties of the detection framework. Spatial explainability was assessed by comparing learned attention or predicted mask support with annotated object-support regions. Predictive reliability was assessed through calibration and Monte Carlo dropout rather than relying on raw confidence values alone.

3.4.1 Quantitative Spatial Localisation

Spatial explainability was evaluated by measuring the overlap between the model-derived support region and the corresponding ground-truth prohibited-item region. For a thresholded attention or predicted mask support region A and the corresponding ground-truth object-support region G , Intersection over Union was defined as

$$\text{IoU}(A, G) = \frac{|A \cap G|}{|A \cup G|}, \quad (21)$$

where A denotes the binary support region obtained from either the thresholded attention map or the predicted mask, G denotes the corresponding ground-truth object-support region, $|A \cap G|$ is

the number of pixels shared by the predicted and ground-truth regions, and $|A \cup G|$ is the total number of pixels belonging to either region. The IoU score ranges from 0 to 1, with larger values indicating stronger spatial agreement between the learned model evidence and the annotated prohibited-item region.

Attention IoU was computed for variants B3–B6, while mask IoU was evaluated only for the mask-guided configurations B5 and B6. Grad-CAM was retained as supplementary post-hoc visual evidence; however, quantitative localisation was emphasized because visually plausible heatmaps alone do not establish that the model is attending to the correct prohibited-item region

3.4.2 Monte Carlo Dropout Uncertainty

For the full B6 configuration, Monte Carlo dropout was used to estimate predictive uncertainty at inference time. Dropout layers within the concealment-refinement module were reactivated while the remainder of the detector remained in evaluation mode. A deterministic forward pass was first used to obtain the reference detections. For each deterministic detection j , a prediction of the same class with an approximately matching bounding box was then sought across T stochastic forward passes. Let $s_j^{(t)}$ denote the confidence score assigned to the matched detection for object j during stochastic pass t . When no suitable match was found in a particular pass, the corresponding confidence score was set to zero.

The mean stochastic confidence for detection j was computed as

$$\bar{s}_j = \frac{1}{T} \sum_{t=1}^T s_j^{(t)}, \quad (22)$$

where $\bar{s}_j$ denotes the mean confidence of detection j across the T Monte Carlo dropout passes, $s_j^{(t)}$ is the matched confidence score from pass t , and T is the total number of stochastic inference passes.

Predictive uncertainty was then quantified using the sample standard deviation of the matched confidence scores:

$$u_j = \sqrt{\frac{1}{T-1} \sum_{t=1}^T (s_j^{(t)} - \bar{s}_j)^2}, \quad (23)$$

where u_j denotes the uncertainty associated with deterministic detection j . A small u_j indicates that the predicted confidence remained stable across stochastic passes, whereas a larger value indicates greater variation in the detector's response and therefore lower predictive stability.

In the final experiment, uncertainty estimation used 12 stochastic forward passes on a deterministic subset of 250 validation images. To assess whether uncertainty was informative for identifying unreliable predictions, detections were labelled according to correctness and the area under the Receiver Operating Characteristic curve was computed with incorrect detections treated as the positive error class. A larger ROC-AUC therefore indicated stronger discrimination between correct and incorrect detections based on the uncertainty score.

3.4.3 Calibration

Calibration analysis treated each retained detection confidence as a probability-like score and defined detection correctness at an IoU threshold of 0.50 as a binary outcome. Temperature scaling was fitted using the validation predictions so that calibration could be adjusted without modifying the detector weights. For a raw confidence score p_i , the corresponding logit was first obtained using the log-odds transformation:

$$z_i = \log\left(\frac{p_i}{1 - p_i}\right), \quad (24)$$

where $p_i \in (0, 1)$ denotes the raw confidence assigned to retained detection i , and z_i is the corresponding unbounded logit representation. In implementation, confidence values at the numerical boundaries were clipped slightly away from 0 and 1 before applying the logarithmic transformation to avoid undefined values.

The logit was then rescaled by a learned temperature parameter $T > 0$, and the calibrated confidence was recovered through the sigmoid function:

$$\hat{p}_i = \sigma\left(\frac{z_i}{T}\right), \quad (25)$$

where T denotes the scalar temperature estimated from the validation detections, $\sigma(\cdot)$ is the sigmoid activation function, and $\hat{p}_i$ is the temperature-scaled confidence for detection i . When $T > 1$, confidence values are generally softened, whereas $T < 1$ produces sharper confidence estimates. Temperature scaling changes the confidence magnitude but does not alter the ranking of detections.

Calibration quality was first quantified using Expected Calibration Error (ECE). The retained detections were partitioned into M confidence bins, and the absolute difference between empirical accuracy and mean confidence was calculated within each bin:

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|, \quad (26)$$

where B_m denotes the set of detections assigned to calibration bin m , $|B_m|$ is the number of detections in that bin, N is the total number of retained detections, $\text{acc}(B_m)$ is the proportion of correct detections in bin m , and $\text{conf}(B_m)$ is the mean predicted confidence within the same bin. An ECE value closer to zero indicates stronger agreement between predicted confidence and observed correctness.

Calibration was additionally assessed using the Brier score, which measures the mean squared difference between predicted confidence and the binary correctness outcome:

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^N (\hat{p}_i - y_i)^2, \quad (27)$$

where $\hat{p}_i$ denotes the evaluated confidence for detection i raw or temperature-scaled depending on the calibration condition and $y_i \in \{0, 1\}$ denotes its binary correctness label, with $y_i = 1$ for a correct detection at $\text{IoU} \geq 0.50$ and $y_i = 0$ otherwise. Lower Brier scores indicate better agreement between predicted confidence and actual detection correctness.

3.5 Experimental Workflow and Evaluation

The final CAFEM experiment used 640×640 images, a training batch size of four, a maximum of eight epochs per enhanced variant, early-stopping patience of two epochs, and separate learning rates for the new module and the fine-tuned Detect head as shown in table 13. The backbone and neck were frozen. Full validation predictions were used to select the operating confidence threshold from a grid spanning 0.10 to 0.90 in steps of 0.05. The held-out Easy, Hard, Hidden, and overall test data were used only after model selection.

Table 13 Principal final CAFEM training and reliability settings

Parameter	Setting
Input resolution	640×640
Training batch size	4
Evaluation batch size	4
Maximum CAFEM epochs	8
Early-stopping patience	2
New-module learning rate	1×10^{-3}
Detect-head learning rate	2×10^{-4}
Weight decay	1×10^{-4}
Validation monitor sample	Up to 400 images
Operating-threshold grid	0.10-0.90 in 0.05 increments
MC-dropout subset	250 validation images
MC-dropout passes	12
External validation	Zero-shot SIXray-derived test subset

3.5.1 Detection Performance Metrics

Detection was evaluated using precision, recall, F1-score, mean Average Precision at IoU 0.50 (mAP@0.50), and COCO-style mAP averaged from IoU 0.50 to 0.95 in increments of 0.05. False positives and false negatives per image were included for operational interpretation. These quantities were computed using the following definitions:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (28)$$

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (29)$$

$$F_1 = 2 \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right), \quad (30)$$

$$AP_{c,\tau} = \int_0^1 P_{c,\tau}(R) dR, \quad (31)$$

$$\text{mAP@0.50} = \frac{1}{C} \sum_{c=1}^C AP_{c,0.50}, \quad (32)$$

$$\text{mAP@0.50:0.95} = \frac{1}{10C} \sum_{\tau \in \{0.50, 0.55, \dots, 0.95\}} \sum_{c=1}^C AP_{c,\tau}, \quad (33)$$

$$\text{FP/image} = \frac{FP}{N_{\text{img}}}, \text{FN/image} = \frac{FN}{N_{\text{img}}}, \quad (34)$$

where TP , FP , and FN denote true-positive, false-positive, and false-negative detections, respectively; C is the number of prohibited-item classes; $AP_{c,\tau}$ is the average precision for class c at IoU threshold τ ; $P_{c,\tau}(R)$ denotes the precision–recall curve for class c at threshold τ ; and N_{img} is the number of evaluated images.

For the repaired final evaluation, COCO-style mAP used the standard maxDets = [1, 10, 100] configuration. This correction was necessary because the earlier [1, 10, 300] setting caused the aggregate mAP@0.50:0.95 field to return the invalid sentinel value -1 even though mAP@0.50 was computed.

3.5.2 Difficulty-Stratified and External Evaluation

Each controlled variant was evaluated separately on the PIDray Easy, Hard, and Hidden test subsets before the subsets were combined for the overall PIDray result. The central concealment-specific hypothesis was therefore evaluated using the official difficulty labels only at test time. The final B6 model was then assessed on the 831-image SIXray-derived common-class subset without fine-tuning and without threshold re-optimization. This final experiment quantified cross-dataset domain transfer rather than within-dataset generalisation.

3.6 Architecture of the Proposed Framework

Figure 15 presents the complete architecture of the proposed CAU-XInspect framework and illustrates the end-to-end flow from X-ray input to detection, explainability, and reliability assessment. The preprocessed X-ray image is first passed through the frozen YOLO26 backbone and neck, which extract hierarchical multi-scale feature representations at different spatial resolutions. These feature maps are then supplied to CAFEM, where each detection scale is independently refined through complementary occlusion, contrast, and contextual branches designed to preserve local object structure, emphasize weak feature differences, and incorporate broader surrounding evidence. The resulting branch representations are fused to generate concealment-aware features and a learned spatial attention map, after which residual enhancement is applied so that the original YOLO26 information is preserved while informative concealment-sensitive features are selectively strengthened.

For the mask-guided configurations, an auxiliary mask branch provides additional spatial supervision and modulates the refined features toward annotated prohibited-item regions. In the full B6 configuration, attention-mask consistency further encourages agreement between the independently learned attention and mask representations. The refined multi-scale features are subsequently passed to the fine-tuned YOLO26 Detect head to generate prohibited-item bounding boxes, class predictions, and confidence scores. Beyond conventional detection, the framework also supports quantitative evaluation of spatial grounding through attention and mask IoU, while calibration analysis and Monte Carlo dropout provide complementary measures of predictive reliability and uncertainty. The architecture therefore integrates concealment-aware feature refinement, spatial supervision, object detection, explainability, and uncertainty assessment within a single prohibited-item X-ray inspection pipeline.

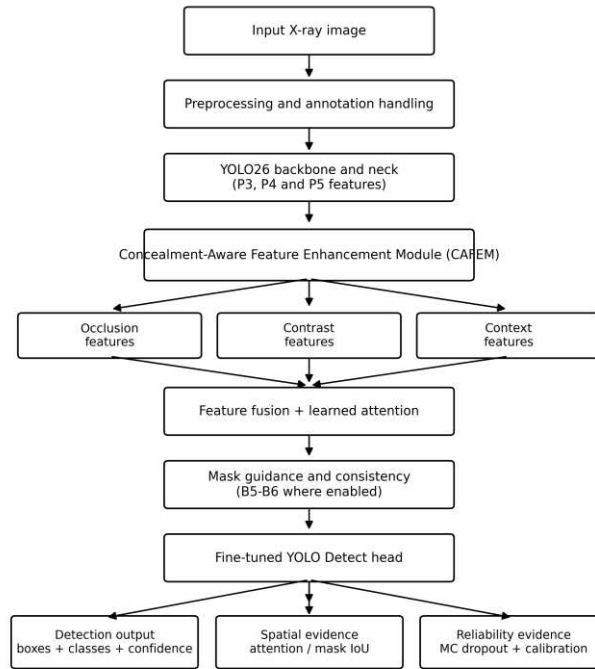


Fig. 15 Architecture of the proposed CAU-XInspect concealment-aware X-ray inspection framework.

3.7 Model Implementation

The framework was implemented using Python 3.12.6, PyTorch 2.8.0+cu129, and Ultralytics 8.4.120, with TorchMetrics, NumPy, Pandas, Matplotlib, OpenCV, and scikit-learn supporting evaluation, data processing, visualization, calibration, and error analysis. Model development, training, and evaluation were conducted in GPU-accelerated cloud environments, including an NVIDIA A10 GPU with 24 GB of GPU memory. Training used separate learning rates of 0.001 for the newly introduced refinement module and 0.0002 for the fine-tuned Detect head, with a weight decay of 0.0001, enabling controlled partial fine-tuning while the YOLO26 backbone and neck remained frozen.

4 Experimental Results

4.1 Overall Performance of the Controlled B1-B6 Models

Table 14 presents the overall performance of the six controlled configurations. The conventional YOLO26 baseline (B1) achieved an $mAP@0.50:0.95$ of 0.5815, $mAP@0.50$ of 0.6973, precision of 0.7854, recall of 0.6106, and F1-score of 0.6870. All enhanced configurations improved $mAP@0.50:0.95$, $mAP@0.50$, and F1-score relative to the baseline, although the magnitude and type of improvement differed across variants.

Table 14 Overall detection performance of the controlled B1-B6 configurations

Variant	Model	Threshold	$mAP@0.50:0.95$	$mAP@0.50$	Precision	Recall	F1-score
B1	YOLO26	0.50	0.5815	0.6973	0.7854	0.6106	0.6870
B2	YOLO26 + CBAM	0.35	0.5946	0.7353	0.8047	0.6948	0.7457
B3	YOLO26 + CAAM	0.35	0.5989	0.7415	0.8029	0.7064	0.7516
B4	YOLO26 + CAFEM	0.40	0.5984	0.7417	0.8303	0.6922	0.7550
B5	YOLO26 + CAFEM + mask guidance	0.40	0.5991	0.7435	0.8305	0.6919	0.7549
B6	YOLO26 + CAFEM + mask guidance + consistency	0.35	0.5968	0.7406	0.8000	0.7111	0.7529

B5 achieved the highest overall $mAP@0.50:0.95$ (0.5991) and $mAP@0.50$ (0.7435), together with the highest precision (0.8305). Relative to B1, B5 improved $mAP@0.50:0.95$ by 0.0177, equivalent to approximately 3.0% relative improvement, and improved $mAP@0.50$ by 0.0461, equivalent to approximately 6.6% relative improvement. B4 produced the highest overall F1-score (0.7550), whereas the full B6 configuration produced the highest recall (0.7111).

The increase in recall for B6 is especially relevant to prohibited-item inspection. Recall rose from 0.6106 for B1 to 0.7111 for B6, an absolute gain of 0.1005 and a relative improvement of approximately 16.5%. This result indicates that the full concealment-aware model recovered a larger proportion of prohibited objects, although its precision of 0.8000 was lower than the precision obtained by B4 and B5. The overall results therefore show a performance trade-off rather than universal dominance by a single configuration: B5 maximized overall AP and precision, B4 maximized F1-score, and B6 maximized sensitivity.

4.2 Ablation Analysis of the Proposed Components

A controlled ablation analysis was conducted to determine the incremental contribution of the successive architectural components relative to the B1 YOLO26 baseline. Because the same underlying detector source, data partitions, and evaluation protocol were retained, the changes observed across B2-B6 provide controlled evidence about how conventional attention, concealment-aware attention, CAFEM, mask guidance, and consistency learning alter detector behavior as shown in table 15 with positive values indicating improvement over B1.

Table 15 Absolute performance change relative to the B1 YOLO26 baseline

Variant	Delta mAP@0.50:0.95	Delta mAP@0.50	Delta F1	Delta Hidden mAP@0.50:0.95	Delta Hidden F1
B2	+0.0131	+0.0380	+0.0587	+0.0346	+0.0745
B3	+0.0174	+0.0442	+0.0645	+0.0353	+0.0839
B4	+0.0169	+0.0444	+0.0680	+0.0324	+0.0868
B5	+0.0177	+0.0461	+0.0678	+0.0323	+0.0862
B6	+0.0154	+0.0432	+0.0659	+0.0325	+0.0878

Adding CBAM in B2 increased overall mAP@0.50:0.95 by 0.0131 and F1-score by 0.0587 relative to B1, showing that feature attention alone produced a measurable gain. Replacing conventional attention with CAAM in B3 increased overall mAP@0.50:0.95 to 0.5989 and produced the largest Hidden mAP@0.50:0.95 improvement over B1 (+0.0353). This result supports the value of attention specifically designed around concealment-related cues rather than generic channel-spatial attention alone.

The transition from CAAM to CAFEM in B4 did not materially increase overall AP, but it raised precision to 0.8303 and produced the highest overall F1-score. Adding mask guidance in B5 yielded the highest overall mAP values and precision. The final consistency term in B6 slightly reduced overall mAP relative to B5 but increased recall to 0.7111 and produced the largest improvement in Hidden F1-score relative to the baseline (+0.0878). Thus, the components influenced different aspects of performance: mask guidance primarily strengthened overall discrimination and precision, whereas the complete consistency-aware configuration shifted the model toward sensitivity and recovery of difficult objects.

4.3 Performance Across PIDray Difficulty Levels

Table 16 reports the $\text{mAP}@0.50:0.95$ values together with Hidden recall and Hidden F1-score. Since concealment, clutter, overlap, and weak object boundaries can affect detection differently from straightforward scenes, aggregate performance alone is insufficient to characterize the proposed framework. Each model was therefore evaluated separately on the PIDray Easy, Hard, and Hidden test subsets.

Table 16 Performance across PIDray difficulty subsets

Variant	Easy $\text{mAP}@0.50:0.95$	Hard $\text{mAP}@0.50:0.95$	Hidden $\text{mAP}@0.50:0.95$	Hidden Recall	Hidden F1
B1	0.6650	0.6217	0.3443	0.3550	0.4600
B2	0.6722	0.6251	0.3789	0.4724	0.5345
B3	0.6784	0.6272	0.3796	0.4766	0.5439
B4	0.6791	0.6273	0.3768	0.4569	0.5468
B5	0.6786	0.6277	0.3766	0.4645	0.5463
B6	0.6753	0.6300	0.3769	0.4918	0.5478

A clear difficulty gradient was observed for all variants. Easy images consistently produced the highest $\text{mAP}@0.50:0.95$, followed by Hard images, whereas Hidden images produced substantially lower values. For B1, $\text{mAP}@0.50:0.95$ decreased from 0.6650 on Easy images to 0.6217 on Hard images and 0.3443 on Hidden images. The same general ordering remained across the enhanced variants, confirming that heavily concealed prohibited items constitute the most challenging condition in the evaluation.

Despite this difficulty, all enhanced models increased Hidden $\text{mAP}@0.50:0.95$ relative to B1. B3 achieved the highest Hidden $\text{mAP}@0.50:0.95$ of 0.3796, an absolute gain of 0.0353 over the baseline. On the Hard subset, B6 achieved the highest $\text{mAP}@0.50:0.95$ of 0.6300, while B4 recorded the highest Easy $\text{mAP}@0.50:0.95$ of 0.6791 as shown in figure 16. These differences indicate that the benefit of each architectural component is condition-dependent rather than uniform across all difficulty levels.

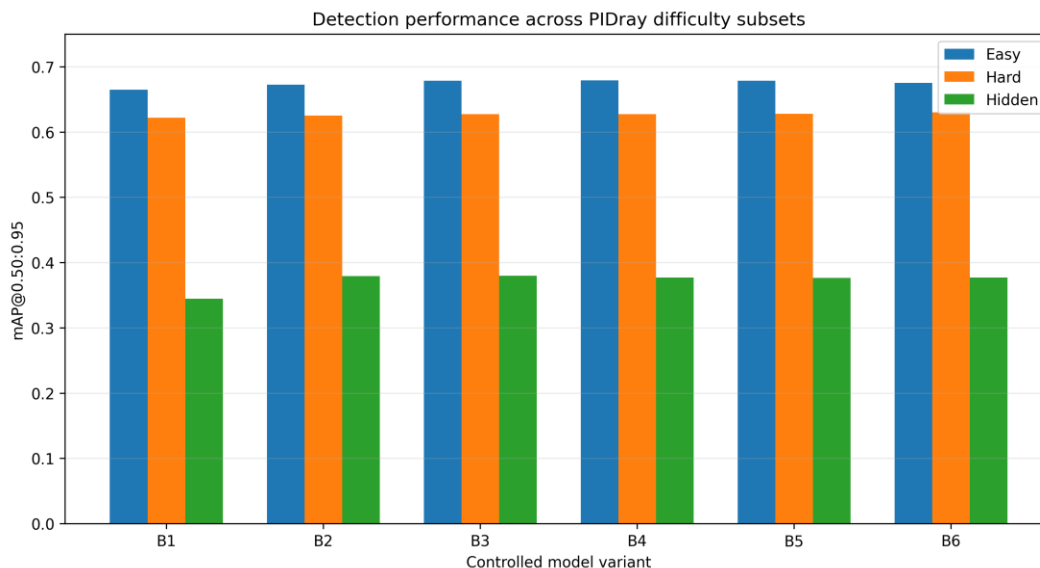


Fig. 16 mAP@0.50:0.95 across Easy, Hard, and Hidden PIDray test subsets.

4.4 Detection Performance on Concealed Prohibited Items

The Hidden subset provides the most direct evaluation of the central concealment-aware objective because these images contain prohibited items for which the available visual evidence is substantially reduced. The conventional YOLO26 detector achieved a Hidden recall of 0.3550 and Hidden F1-score of 0.4600. The enhanced configurations consistently improved these operational measures. The full B6 model achieved a Hidden recall of 0.4918, compared with 0.3550 for B1 as illustrated in figure 17. This represents an absolute gain of 0.1368 and a relative improvement of approximately 38.5%. Hidden F1-score similarly increased from 0.4600 to 0.5478, corresponding to an absolute gain of 0.0878 and a relative improvement of approximately 19.1%. B6 therefore achieved the strongest Hidden recall and Hidden F1-score among the evaluated variants.

Importantly, the highest Hidden mAP@0.50:0.95 was obtained by B3 rather than B6. This distinction separates average-precision ranking from operational sensitivity at the validation-selected threshold. B3 yielded the strongest Hidden AP, whereas B6 recovered the largest proportion of concealed items and achieved the best balance between Hidden precision and recall as reflected by the highest Hidden F1-score. The result therefore supports a sensitivity-oriented interpretation of the full model rather than a claim that B6 dominates every metric.

The overall error analysis reinforces this interpretation. Relative to B1, B6 reduced false negatives per image by 0.1290, the largest reduction among the controlled variants, while false positives per image increased by 0.0139. By contrast, B4 and B5 reduced both false positives and false negatives relative to B1. Consequently, B6 should be interpreted as the most sensitivity-oriented configuration, whereas B4 and B5 offer a more conservative precision-oriented operating profile.

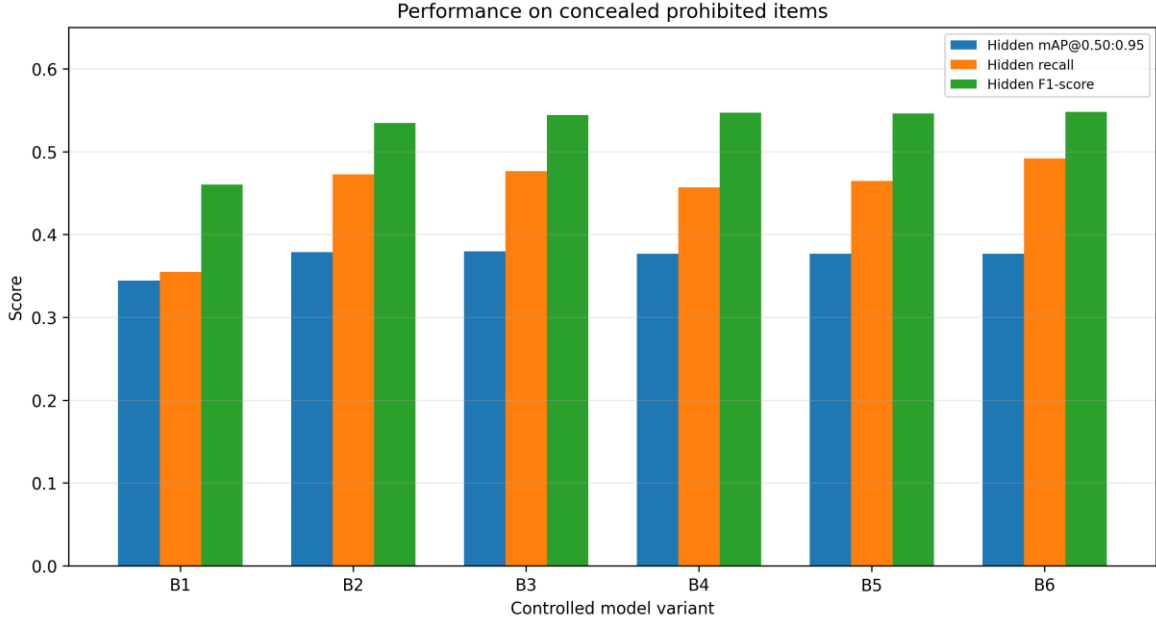


Fig. 17 Hidden-item mAP@0.50:0.95, recall, and F1-score across B1-B6.

4.5 Learned Localization and Explainability Performance

Detection performance was complemented by quantitative localization analysis to determine whether the learned attention increasingly concentrated on annotated prohibited-item regions. This is important because a visually plausible attention map does not, by itself, establish that the model is using relevant spatial evidence. Attention Intersection over Union (IoU) was therefore used to quantify overlap between the learned attention representation and object regions, while the mask-guided variants were additionally assessed using mask IoU. Table 17 presents the resulting localization performance for variants B3–B6, showing the degree of spatial agreement achieved by the attention and mask-guided components.

Table 17 Validation localisation performance.

Variant	Attention IoU	Mask IoU
B3	0.6815	—
B4	0.7087	—
B5	0.7447	0.7383
B6	0.7578	0.7777

Attention IoU increased progressively from 0.6815 for B3 to 0.7087 for B4, 0.7447 for B5, and 0.7578 for B6. This monotonic improvement is notable because overall detection AP did not increase monotonically across the same variants. The localization results therefore provide independent evidence that the later concealment-aware components improved spatial grounding even when their contribution to aggregate AP was comparatively modest.

Mask localization showed the same direction of improvement. B5 achieved a mask IoU of 0.7383, while B6 increased mask IoU to 0.7777. The full configuration therefore produced the strongest correspondence between learned spatial emphasis and annotated prohibited-item regions. Taken together, these results indicate that CAFEM and the mask/consistency extensions did more than alter detection scores: they progressively strengthened the spatial alignment of the learned representation with the relevant object regions.

4.6 Predictive Reliability, Calibration, and Uncertainty

The final B6 model was further evaluated to determine whether its confidence values and stochastic uncertainty estimates could provide useful information about prediction reliability. Calibration was assessed using Expected Calibration Error (ECE) and Brier score, while error discrimination was evaluated using the area under the receiver operating characteristic curve (AUC) for confidence- and uncertainty-based separation of correct and incorrect predictions. Table 18 summarizes the resulting calibration and uncertainty-reliability measures for the B6 model, including the raw and temperature-scaled calibration results and the error-discrimination performance obtained from deterministic confidence and Monte Carlo dropout uncertainty.

Table 18 Reliability and uncertainty results for B6

Measure	Result
Raw ECE	0.0095
Temperature-scaled ECE	0.0098
Raw Brier score	0.0146
Temperature-scaled Brier score	0.0143
Inverse-confidence error AUC	0.9894
MC-dropout error AUC	0.8873
MC-dropout evaluation images	250
Stochastic forward passes	12

The raw detector achieved an ECE of 0.0095 and a Brier score of 0.0146. After temperature scaling, ECE was 0.0098 and the Brier score was 0.0143. The very small raw ECE shows that the detector was already well calibrated before post-hoc temperature scaling. Temperature scaling therefore did not improve ECE, however, it produced a small improvement in the Brier score. The Inverse-confidence error AUC was 0.9894, showing very strong separation between reliable and erroneous predictions using detector confidence. Figure 18 compares the calibration behaviour of the B6 model before and after temperature scaling

Monte Carlo dropout was evaluated on 250 validation images using 12 stochastic forward passes and produced 309 evaluated predictions. Mean uncertainty was 0.00687, while the uncertainty-based error-discrimination AUC was 0.8873. Because an AUC of 0.5 would indicate chance-level discrimination, the observed value provides strong subset-level evidence that higher uncertainty was associated with prediction errors. This uncertainty result complements the detector-confidence analysis by showing that stochastic variation in the model output also carries information about likely failures.

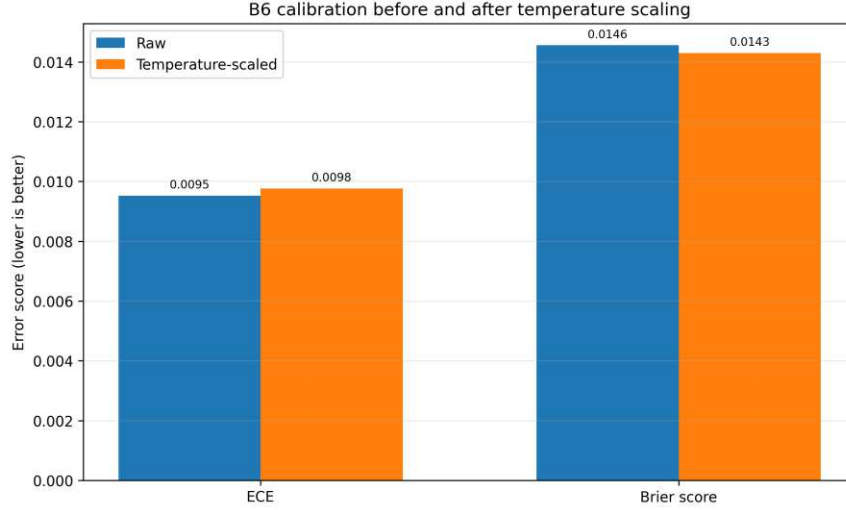


Fig. 18 B6 calibration before and after temperature scaling

4.7 Zero-Shot External Validation on the SIXray-Derived Dataset

Cross-dataset robustness was evaluated by applying the final B6 model to an independent SIXray-derived test set without fine-tuning. The external evaluation comprised 831 images and was restricted to the five prohibited-item categories shared semantically between PIDray and the SIXray-derived dataset: Gun, Knife, Pliers, Scissors, and Wrench. Class identities were explicitly mapped between the datasets, and the operating threshold was retained from PIDray validation rather than optimized on SIXray. This preserved the zero-shot nature of the experiment.

External performance was substantially lower than the internal PIDray performance. B6 achieved an internal PIDray $\text{mAP}@0.50:0.95$ of 0.5968 and $\text{mAP}@0.50$ of 0.7406, whereas zero-shot SIXray performance decreased to 0.0408 and 0.0962, respectively. Recall similarly decreased from 0.7111 on PIDray to 0.0954 on the SIXray-derived test set. Precision on the external set was 0.3918 and F1-score was 0.1535 as shown in table 19.

Table 19 Zero-shot B6 performance on the SIXray-derived external test set

Metric	Result
$\text{mAP}@0.50:0.95$	0.0408
$\text{mAP}@0.50$	0.0962
Precision	0.3918
Recall	0.0954
F1-score	0.1535
Number of test images	831

The external experiment therefore does not demonstrate strong zero-shot generalization across the two X-ray datasets as shown in figure 19. Instead, the pronounced reduction in detection performance provides quantitative evidence of substantial cross-dataset domain shift. The finding is scientifically important because it distinguishes strong within-domain detection and concealment sensitivity from robustness to differences in scanner characteristics, acquisition

conditions, background distributions, image statistics, and annotation conventions. The SIXray result should therefore be interpreted as an identified deployment limitation and a motivation for future domain adaptation or domain-generalisation research, rather than being obscured by reporting only the stronger PIDray results.

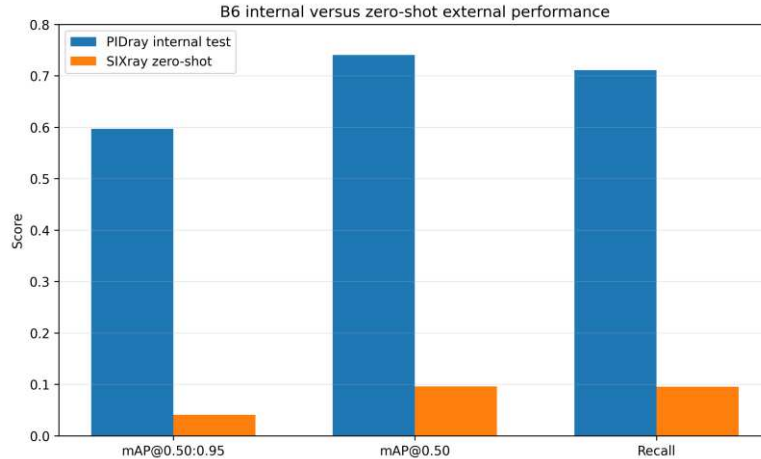


Fig. 19 Internal PIDray versus zero-shot SIXray performance of the full B6 model.

5 Discussion of Results

5.1 Overall Detection Performance and the Contribution of the Ablation Sequence

The central result from the controlled ablation showed that the concealment-aware extensions improved the YOLO26 baseline, but they did not produce a single configuration that was best on every metric. B5 achieved the highest overall mAP@0.50:0.95 (0.5991), mAP@0.50 (0.7435), and precision (0.8305); B4 achieved the highest F1-score (0.7550); and the full B6 model achieved the highest recall (0.7111). This pattern is more informative than a simple ranking because it shows that the added modules altered the operating characteristics of the detector rather than merely shifting all metrics upward by the same amount.

The improvement from B1 to B2 indicates that conventional attention already provides useful feature reweighting in X-ray imagery. The subsequent move from CBAM to CAAM and CAFEM produced further gains, supporting the idea that attention becomes more useful when it is adapted to the structural difficulties of baggage X-rays rather than applied as a generic visual module. This interpretation is consistent with earlier prohibited-item research. [8] designed PIDray and the Selective Dense Attention Network specifically around intra-class variation, class imbalance, occlusion, and deliberately hidden objects. Similarly, [9] reported that a de-occlusion attention mechanism produced its clearest benefits under stronger occlusion on OPIXray. [7] likewise argued that spatial structure, contour information, and multi-scale global context are important for detecting occluded prohibited objects. The present findings therefore converge with the literature at the mechanism level where targeted feature enhancement is most useful when conventional detector features are degraded by overlap, clutter, weak boundaries, or concealment.

Adding consistency learning in B6 did not improve overall mAP beyond B5 instead, it shifted the model toward greater sensitivity. This suggests that the final consistency objective encouraged the model to retain or recover more candidate prohibited items, including ambiguous ones, but at the cost of some precision. In other words, the final model should not be interpreted as a universally superior detector. It is better understood as a sensitivity-oriented configuration within the same family, whereas B5 represents the strongest accuracy-and-precision configuration. This distinction is particularly relevant in security inspection, where the relative cost of a missed prohibited item and an additional false alarm determine which, operating point is preferable.

The ablation therefore supports the architectural contribution without requiring an exaggerated claim of universal superiority. The study shows that conventional attention, concealment-aware attention, CAFEM, mask guidance, and consistency learning contribute different benefits. This is consistent with the broader studies where complementary modules should be judged by the specific function they add to the pipeline rather than by assuming that more components must always yield the single highest aggregate score[42][43].

5.2 Concealment-Specific Performance and the Significance of Hidden-Item Recall

The Easy-Hard-Hidden analysis provides the strongest evidence that the study addressed. Across all models, performance decreased as the task moved from Easy to Hard and then to Hidden images. For B1, $\text{mAP}@0.50:0.95$ fell from 0.6650 on Easy images to 0.6217 on Hard images and 0.3443 on Hidden images. This confirms that concealment remains a major source of detection difficulty even for a strong modern detector. The pattern agrees with the motivation of PIDray, which was explicitly developed to include deliberately hidden prohibited items [8], and with OPIXray, where increasing occlusion was shown to degrade ordinary detector performance and to increase the relative value of de-occlusion attention [9].

The most operationally important improvement occurred in Hidden recall. B6 increased Hidden recall from 0.3550 to 0.4918, an absolute gain of 0.1368 and a relative improvement of approximately 38.5%. Hidden F1-score increased from 0.4600 to 0.5478. These gains are more important to the central research problem than the relatively modest change in overall $\text{mAP}@0.50:0.95$ because hidden prohibited items are precisely the cases for which ordinary X-ray detection is most likely to fail. The result suggests that concealment-aware enhancement recovered visual evidence that the baseline either suppressed or failed to exploit sufficiently.

However, the Hidden results also demonstrate why average precision, recall, precision, and F1-score should be interpreted together. B3 achieved the highest Hidden $\text{mAP}@0.50:0.95$ (0.3796), whereas B6 achieved the highest Hidden recall and F1-score. The difference indicates that B3 ranked detections most favourably across confidence thresholds, while B6 was more effective at recovering hidden items at the selected operational threshold. Consequently, the study does not support the claim that B6 is the best hidden-item detector under every criterion. Rather, it supports the narrower and more defensible claim that the full configuration gives the strongest sensitivity to hidden items at the chosen operating point.

This sensitivity is accompanied by a false-positive trade-off. The earlier analysis showed that B6 reduced Hidden false negatives per image relative to B1 but increased Hidden false positives. In a security-screening workflow, such a trade-off may still be acceptable when missing a prohibited item carries a larger consequence than sending an additional bag for secondary inspection. Nevertheless, the preferred threshold should ultimately be selected using operational cost, inspection workload, prevalence, and human-review capacity rather than detector metrics alone. This is why the present model should be viewed as a decision-support component rather than an autonomous screening authority.

5.3 Explainability and Spatial Grounding of Concealment-Aware Features

Attention IoU increased progressively from 0.6815 for B3 to 0.7087 for B4, 0.7447 for B5, and 0.7578 for B6. Mask IoU also increased from 0.7383 in B5 to 0.7777 in B6. This progression is important because overall detection mAP did not rise monotonically across the same sequence. Therefore, the later modules appear to improve where the model concentrates spatial evidence even when they do not produce the highest aggregate detection score.

This distinction strengthens the scientific contribution. A heatmap that appears visually plausible is not, by itself, evidence that the detector relies on the correct object region. Grad-CAM was originally proposed as a class-discriminative localization method that highlights image regions associated with a target prediction [30], but visual inspection can still be subjective. The present study therefore follows a more rigorous principle where explanation or attention maps are compared with annotated object regions using IoU. A similar argument by [44], also treated quantitative overlap with lesion-support regions as stronger evidence than attractive saliency maps alone.

The monotonic localization improvement from B3 to B6 suggests that CAFEM, mask guidance, and consistency learning progressively constrain the representation toward the actual prohibited-item region. This is consistent with the intent of the architecture where the mask branch provides explicit spatial supervision, while the consistency term discourages disagreement between the enhanced attention representation and object evidence. The B6 mask IoU of 0.7777 is therefore not merely an auxiliary metric; it shows that the full configuration achieved the strongest spatial correspondence with the annotated prohibited-object regions among the tested variants.

Nevertheless, this evidence should be described as spatial plausibility or localization consistency rather than proof of causal faithfulness. High overlap shows that the learned attention corresponds to the object region, but it does not establish that every highlighted pixel was necessary for the detection decision or that the model is free from reliance on contextual cues. A future explanation study could combine localization IoU with perturbation-based faithfulness tests, explanation stability across augmentations, and inspector assessment of whether highlighted regions are operationally meaningful.

5.4 Predictive Uncertainty, Calibration, and Reliability

The uncertainty results add a reliability layer that is separate from detection accuracy. The raw B6 detector achieved an Expected Calibration Error of 0.0095 and a Brier score of 0.0146, indicating close agreement between reported confidence and observed correctness on the evaluated validation data. Temperature scaling changed ECE only marginally to 0.0098 while slightly improving the Brier score to 0.0143. This is an important result because temperature scaling is widely used as a post-hoc calibration method [35], yet the present model was already strongly calibrated under the reported evaluation protocol. Temperature scaling did not improve calibration overall; it produced a small Brier-score improvement without improving ECE.

Monte Carlo dropout provided an additional measure of uncertainty. The error-discrimination AUC of 0.8873 on the 250-image validation subset indicates that uncertainty was substantially higher for many erroneous predictions than for reliable predictions. This is consistent with the use of dropout as an approximate Bayesian inference mechanism for representing predictive uncertainty [34]. The result is also conceptually aligned with study by [44] where uncertainty and calibration were evaluated as a safety layer rather than as substitutes for predictive performance.

The distinction between calibration and error discrimination is important. Calibration asks whether a stated probability corresponds to empirical correctness frequency, whereas error-discrimination AUC asks whether a reliability score can rank incorrect predictions above correct ones. The very low ECE and strong Inverse-confidence error AUC indicate that the B6 detector provides useful confidence information, while the MC-dropout AUC suggests that stochastic uncertainty adds additional error-sensitive evidence. However, the MC-dropout analysis was limited to 250 images and 309 evaluated predictions. It should therefore be described as promising subset-level evidence rather than a definitive estimate of uncertainty performance across the entire test distribution.

5.5 Cross-Dataset Generalization and the PIDray-to-SIXray Domain Shift

The zero-shot SIXray experiment is the clearest boundary on the generalizability of the proposed framework. B6 achieved $\text{mAP}@0.50:0.95$ of 0.5968 and $\text{mAP}@0.50$ of 0.7406 on the internal PIDray evaluation, but only 0.0408 and 0.0962, respectively, on the 831-image SIXray-derived common-class test set. Recall fell from 0.7111 to 0.0954. The size of this reduction means that the model cannot be described as cross-dataset robust in its current form.

This result is nevertheless scientifically valuable. SIXray was designed around large-scale security inspection with severe object overlap, complex backgrounds, and extreme class imbalance [10], whereas PIDray emphasizes a different distribution of real-world prohibited items, including deliberately hidden examples and segmentation annotations [8]. Differences in scanner characteristics, material colour mapping, object prevalence, acquisition conditions, class composition, image resolution, background clutter, and annotation practice can all alter the feature distribution available to a detector. The observed performance collapse is therefore consistent with a substantial dataset shift, although the present experiment does not isolate which source of shift is dominant.

The external experiment also strengthens the validity of the study because it prevents within-dataset performance from being mistaken for deployment generalizability. [43] for example, explicitly identifies the absence of external validation as a limitation and recommends evaluation on unseen datasets from different acquisition environments. In the present study, external validation was actually performed, and it revealed a severe limitation that would otherwise have remained hidden. This is a more informative outcome than reporting only strong internal results.

The SIXray result therefore changes the deployment interpretation of CAU-XInspect. The framework can be defended as effective within the controlled PIDray setting and as sensitive to deliberately hidden objects, but not yet as a scanner-independent prohibited-item detector. Future work should investigate domain generalization, target-domain calibration, style-robust feature learning, scanner-specific normalization, unsupervised or semi-supervised domain adaptation, and multi-dataset training. Any such extension should preserve a held-out external dataset so that adaptation gains are not evaluated on data used to tune the model.

5.6 Practical Implications for Non-Intrusive X-Ray Inspection

From an operational perspective, the results support a human-in-the-loop interpretation of the framework. B5 and B6 offer two useful but different operating profiles. B5 is the stronger overall AP/precision configuration, while B6 is the stronger sensitivity configuration, particularly for hidden items. This means deployment should not begin by declaring one model universally best; instead, an inspection authority should define the relative cost of missed prohibited items and false alarms and then select the model configuration and threshold that match those costs.

In high-security contexts, B6 may be preferable because its higher recall and lower hidden-item false-negative burden reduce the chance that a concealed item is missed. However, the increase in false positives would increase secondary-inspection workload. In lower-prevalence or high-throughput environments, B5 may provide a more practical balance because its higher precision and overall AP can reduce unnecessary interventions. The model therefore has potential value as a configurable decision-support system rather than a fixed one-size-fits-all detector.

The explainability and uncertainty layers also have practical relevance. Spatial localization can show where the model concentrated evidence, while calibrated confidence and stochastic uncertainty can identify predictions that deserve closer scrutiny. These functions mirror a broader principle in safety-critical AI where reliability depends not only on whether a model is accurate, but also on whether its evidence is interpretable and whether low-confidence cases can be recognized. Despite these advantages, the present results do not justify autonomous deployment. No human-inspector study was conducted, no prospective screening workflow was evaluated, and the severe SIXray domain shift shows that performance may deteriorate sharply outside the development distribution. Before real-world use, the system would require local validation on the intended scanner fleet, target-site threshold selection, calibration monitoring, workload analysis, false-alarm review, cybersecurity and data-governance controls, and human-factors testing with trained inspectors.

5.7 Limitations and Future work.

First, the study was evaluated primarily on publicly available X-ray security datasets, with PIDray serving as the principal development and internal evaluation dataset. Although the official Easy, Hard, and Hidden subsets provided controlled assessment across different levels of detection difficulty, the findings may not fully represent variations arising from different scanner manufacturers, acquisition protocols, baggage compositions, and deployment environments. Future work should therefore evaluate CAU-XInspect on additional independently collected security X-ray datasets and, where possible, on images acquired from operational screening environments to establish broader cross-scanner and cross-site generalizability.

Second, external validation was restricted to an 831-image SIXray-derived subset containing five classes that were semantically shared with PIDray: Gun, Knife, Pliers, Scissors, and Wrench. This restriction ensured a valid common-label comparison but does not establish performance across the complete SIXray label space or other prohibited-item categories. The observed reduction in zero-shot performance further demonstrates the effect of domain shift between X-ray datasets. Future work should investigate multi-dataset training, domain-generalization strategies, domain adaptation, and scanner-specific recalibration to improve robustness across heterogeneous security-imaging environments.

Third, the study evaluated explainability and predictive reliability using quantitative computational measures but did not assess how these outputs influence human security-screening decisions. Attention IoU, mask IoU, calibration, and uncertainty analysis demonstrate that the framework can provide spatial and reliability information; however, their practical usefulness to security personnel remains to be established. Future work should therefore conduct human-in-the-loop studies involving trained X-ray screeners to determine whether attention maps, mask guidance, confidence scores, and uncertainty indicators improve detection accuracy, decision time, and user trust during realistic screening tasks.

Conclusion

This study developed CAU-XInspect, a concealment-aware, explainable, and uncertainty-aware framework for prohibited-item detection in security X-ray images. The proposed CAFEM-based variants improved the controlled YOLO26 baseline, with B5 achieving the highest overall $mAP@0.50:0.95$ of 0.5991, while the full B6 configuration achieved the highest recall of 0.7111 and the strongest sensitivity to concealed prohibited items. On the Hidden subset, B6 increased recall from 0.3550 to 0.4918 and F1-score from 0.4600 to 0.5478, demonstrating the value of concealment-aware feature refinement for difficult X-ray cases. Spatial explainability also improved progressively, with attention IoU reaching 0.7578 and mask IoU 0.7777 in B6, indicating stronger alignment between learned model evidence and annotated prohibited-item regions. Reliability analysis further showed that detector confidence was already well calibrated, while Monte Carlo dropout uncertainty provided useful discrimination between correct and incorrect predictions.

Collectively, these findings demonstrate that prohibited-item X-ray detection should be evaluated as an integrated framework encompassing detection accuracy, concealment sensitivity, spatial explainability, calibration, and predictive uncertainty, rather than solely as a conventional object-detection task. However, the marked reduction in zero-shot performance on the independent SIXray-derived subset shows that strong internal PIDray performance does not necessarily translate into robust cross-dataset generalization. The present findings therefore do not support autonomous operational deployment of CAU-XInspect in real-world security screening environments. Instead, the framework should be regarded as a technical proof of concept and a methodological blueprint for future concealment-aware X-ray inspection systems. Further work should include broader multi-dataset and cross-scanner validation, domain-generalization or adaptation strategies, evaluation on locally acquired operational X-ray data, and human-in-the-loop studies with trained security screeners to determine whether the framework improves detection performance, decision confidence, and screening efficiency.

Declarations

Authors contribution

Paul Kategaya Kisozi: Conceptualization, methodology, investigation, original draft, review and editing.

Data availability

All data used in this study were derived from publicly available security X-ray datasets described in the manuscript. PIDray served as the primary development and internal evaluation dataset, including its official Easy, Hard, and Hidden subsets, while an independent SIXray-derived common-class subset was used for zero-shot external validation. The PIDray dataset is publicly available at <https://github.com/bywang2018/security-dataset> and the SIXray dataset is publicly available at <https://github.com/MeioJane/SIXray>. Full bibliographic details for the dataset sources are provided in the reference list [8], [10].

Funding statement: The author received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors

Ethics Approval: Not applicable.

Consent to publish: Not applicable.

Consent to participate: Not applicable

Competing interests: The authors declare no competing interests.

REFERENCES

- [1] N. Jaccard, T. W. Rogers, E. J. Morton, and L. D. Griffin, "Tackling the x-ray cargo inspection challenge using machine learning," *Anom. Detect. Imaging with X-Rays*, vol. 9847, no. i, p. 98470N, 2016, doi: 10.1117/12.2222765.
- [2] J. Cani *et al.*, "Illicit object detection in X-ray imaging using deep learning techniques: A comparative evaluation," 2025, [Online]. Available: <http://arxiv.org/abs/2507.17508>
- [3] Z.-W. Wang, Y. Shen, W. Li, and Z. Yuan, "Prohibited Item Detection in X-ray images based on refined surface perception," *J. Vis. Commun. Image Represent.*, vol. 119, p. 104864, 2026, doi: <https://doi.org/10.1016/j.jvcir.2026.104864>.
- [4] Z. Zhu, Y. Zhu, H. Wang, N. Wang, J. Ye, and X. Ling, "FDTNet: Enhancing frequency-aware representation for prohibited object detection from X-ray images via dual-stream transformers," *Eng. Appl. Artif. Intell.*, vol. 133, p. 108076, 2024, doi: <https://doi.org/10.1016/j.engappai.2024.108076>.
- [5] F. Shao, J. Liu, P. Wu, Z. Yang, and Z. Wu, "Exploiting foreground and background separation for prohibited item detection in overlapping X-Ray images," *Pattern Recognit.*, vol. 122, p. 108261, 2022, doi: <https://doi.org/10.1016/j.patcog.2021.108261>.
- [6] B. Wang, Y. Tian, J. Wang, J. Hu, D. Liu, and Z. Xu, "Detect occluded items in X-ray baggage inspection," *Comput. Graph.*, vol. 115, pp. 148–157, 2023, doi: <https://doi.org/10.1016/j.cag.2023.07.013>.
- [7] C. Ma, L. Zhuo, J. Li, Y. Zhang, and J. Zhang, "Occluded prohibited object detection in X-ray images with global Context-aware Multi-Scale feature Aggregation," *Neurocomputing*, vol. 519, pp. 1–16, 2023, doi: <https://doi.org/10.1016/j.neucom.2022.11.034>.
- [8] B. Wang, L. Zhang, L. Wen, X. Liu, and Y. Wu, "Towards Real-World Prohibited Item Detection : A Large-Scale X-ray Benchmark," pp. 5412–5421.
- [9] Y. Wei, R. Tao, Z. Wu, Y. Ma, L. Zhang, and X. Liu, "Occluded Prohibited Items Detection: An X-ray Security Inspection Benchmark and De-occlusion Attention Module," *MM 2020 - Proc. 28th ACM Int. Conf. Multimed.*, pp. 138–146, 2020, doi: 10.1145/3394171.3413828.
- [10] C. Miao *et al.*, "SIXray: A Large-Scale Security Inspection X-Ray Benchmark for Prohibited Item Discovery in Overlapping Images," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2114–2123. doi: 10.1109/CVPR.2019.00222.
- [11] M. Li *et al.*, "AO-DETR: Anti-Overlapping DETR for X-Ray Prohibited Items Detection.," *IEEE Trans. neural networks Learn. Syst.*, vol. 36, no. 7, pp. 12076–12090, Jul. 2025, doi: 10.1109/TNNLS.2024.3487833.
- [12] X. Meng, H. Feng, Y. Ren, H. Zhang, W. Zou, and X. Ouyang, "Transformer-based dual-view X-ray security inspection image analysis," *Eng. Appl. Artif. Intell.*, vol. 138, p. 109382, 2024, doi: <https://doi.org/10.1016/j.engappai.2024.109382>.
- [13] P. Pai, S. K. Kumar, and G. A. Kumar, "Improving Prohibited Item Detection in X-Ray Images Using Neural Networks Under Complex Occlusion Conditions," *IEEE Access*, vol.

- 13, pp. 182608–182620, 2025, doi: 10.1109/ACCESS.2025.3625103.
- [14] B. Pathiraja, M. Gunawardhana, and M. H. Khan, “Multiclass Confidence and Localization Calibration for Object Detection,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2023-June, pp. 19734–19743, 2023, doi: 10.1109/CVPR52729.2023.01890.
 - [15] T. Popordanoska, A. Tiulpin, and M. B. Blaschko, “Beyond Classification: Definition and Density-based Estimation of Calibration in Object Detection,” *Proc. - 2024 IEEE Winter Conf. Appl. Comput. Vision, WACV 2024*, pp. 574–583, 2024, doi: 10.1109/WACV57701.2024.00064.
 - [16] D. Huseljic, M. Herde, P. Hahn, M. Mjude, and B. Sick, “Systematic Evaluation of Uncertainty Calibration in Pretrained Object Detectors,” *Int. J. Comput. Vis.*, vol. 133, no. 3, pp. 1033–1047, 2025, doi: 10.1007/s11263-024-02219-z.
 - [17] X. Zhang, J. Qiao, X. Wang, W. Zhang, J. Li, and Y. Shi, “PDSXray: A Benchmark Dataset for Pseudocoloring-Driven Domain Adaptation in Security X-ray Inspection,” *Sci. Data*, vol. 13, no. 1, p. 817, 2026, doi: 10.1038/s41597-026-07149-8.
 - [18] F. Yang, R. Jiang, Y. Yan, J.-H. Xue, B. Wang, and H. Wang, “Dual-Mode Learning for Multi-Dataset X-Ray Security Image Detection,” *IEEE Trans. Inf. Forensics Secur.*, vol. 19, pp. 3510–3524, 2024, doi: 10.1109/TIFS.2024.3364368.
 - [19] T. Hassan *et al.*, “Incremental convolutional transformer for baggage threat detection,” *Pattern Recognit.*, vol. 153, p. 110493, 2024, doi: <https://doi.org/10.1016/j.patcog.2024.110493>.
 - [20] H. Zhang, Z. Zhao, and J. Yang, “Attention-based prohibited item detection in X-ray images during security checking,” *IET Image Process.*, vol. 18, no. 5, pp. 1119–1131, Apr. 2024, doi: <https://doi.org/10.1049/ipr2.13013>.
 - [21] J. Park, G. An, B.-N. Lee, and H. Seo, “Real-time CNN-based object detection of prohibited items for X-ray security screening,” *Radiat. Phys. Chem.*, vol. 232, p. 112681, 2025, doi: <https://doi.org/10.1016/j.radphyschem.2025.112681>.
 - [22] H. Zhang, W. Teng, X. He, H. Que, and Y. Zhang, “Lightweight prohibited items detection model in X-ray images based on improved YOLOv7-tiny,” *J. Franklin Inst.*, vol. 362, no. 1, p. 107421, 2025, doi: <https://doi.org/10.1016/j.jfranklin.2024.107421>.
 - [23] Y. Ren *et al.*, “Feature knowledge distillation-based model lightweight for prohibited item detection in X-ray security inspection images,” *Adv. Eng. Informatics*, vol. 65, p. 103125, 2025, doi: <https://doi.org/10.1016/j.aei.2025.103125>.
 - [24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “‘Grad-cam: Visual explanations from deep networks via gradient-based localization.’ Proceedings of the IEEE international conference on computer vision,” 2017, [Online]. Available: <http://gradcam.cloudcv.org>
 - [25] C. Qiu, F. Jin, and Y. Zhang, “Empowering CAM-Based Methods with Capability to Generate Fine-Grained and High-Faithfulness Explanations,” *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 5, pp. 4587–4595, 2024, doi: 10.1609/aaai.v38i5.28258.

- [26] V. Buono, P. S. Mashhadi, M. Rahat, P. Tiwari, and S. Byttner, “Expected Grad-CAM: Towards gradient faithfulness,” 2024, [Online]. Available: <https://arxiv.org/abs/2406.01274v2>
- [27] S. Rahimiaghdam and H. Alemdar, “Evaluating the quality of visual explanations on chest X-ray images for thorax diseases classification,” *Neural Comput. Appl.*, vol. 36, no. 17, pp. 10239–10255, 2024, doi: 10.1007/s00521-024-09587-0.
- [28] A. A. Saeed and R. D. Rashid, “Beyond Black-Box AI: A Quantitative Grad-CAM Analysis of Convolutional Neural Network Interpretability in COVID-19 Chest X-Ray Classification,” *Zanco J. Pure Appl. Sci.*, vol. 38, no. 2, pp. 183–196, 2026, doi: 10.21271/ZJPAS.38.2.13.
- [29] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, and A. Preece, “Sanity checks for saliency metrics,” *AAAI 2020 - 34th AAAI Conf. Artif. Intell.*, no. NeurIPS, pp. 6021–6029, 2020, doi: 10.1609/aaai.v34i04.6064.
- [30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626. doi: 10.1109/ICCV.2017.74.
- [31] Y. Li, C. Zhang, S. Sun, and G. Yang, “X-ray Detection of Prohibited Item Method Based on Dual Attention Mechanism,” *Electron.*, vol. 12, no. 18, 2023, doi: 10.3390/electronics12183934.
- [32] Mingyuan Li, Bowen Ma, Tong Jia, and Yichun Zhang, “PIXDet: Prohibited Items X-Ray Image Detection in Complex Background,” Antonio J Tallón-Ballesteros, Ed., SAGE Publications, 2022. doi: 10.3233/FAIA220521 BT - Proceedings of CECNet 2022.
- [33] F. Zhu, X.-Y. Zhang, Z. Cheng, and C.-L. Liu, “Revisiting Confidence Estimation: Towards Reliable Failure Prediction,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 3370–3387, 2024, doi: 10.1109/TPAMI.2023.3342285.
- [34] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning,” in *Proceedings of The 33rd International Conference on Machine Learning*, M. F. Balcan and K. Q. Weinberger, Eds., in Proceedings of Machine Learning Research, vol. 48. New York, New York, USA: PMLR, 2016, pp. 1050–1059. [Online]. Available: <https://proceedings.mlr.press/v48/gal16.html>
- [35] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On Calibration of Modern Neural Networks,” in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., in Proceedings of Machine Learning Research, vol. 70. PMLR, 2017, pp. 1321–1330. [Online]. Available: <https://proceedings.mlr.press/v70/guo17a.html>
- [36] T. Dimitriadis, T. Gneiting, and A. I. Jordan, “Evaluating probabilistic classifiers: Reliability diagrams and score decompositions revisited,” pp. 1–16, 2020, doi: 10.1073/pnas.2016191118.
- [37] T. Dimitriadis, T. Gneiting, and A. I. Jordan, “Stable reliability diagrams for probabilistic classifiers,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 118, no. 8, Feb. 2021, doi:

10.1073/pnas.2016191118.

- [38] T. Gneiting, F. Balabdaoui, and A. E. Raftery, “Probabilistic Forecasts, Calibration and Sharpness,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 69, no. 2, pp. 243–268, Apr. 2007, doi: 10.1111/j.1467-9868.2007.00587.x.
- [39] X. Huang and Y. Zhang, “ScanGuard-YOLO: Enhancing X-ray Prohibited Item Detection,” 2024.
- [40] K. Zhao, S. Peng, Y. Li, and T. Lu, “A lightweight Xray-YOLO-Mamba model for prohibited item detection in X-ray images using selective state space models,” *Sci. Rep.*, vol. 15, no. 1, p. 13171, 2025, doi: 10.1038/s41598-025-96035-1.
- [41] S. Lin, T. Jia, H. Wang, B. Ma, M. Li, and D. Chen, “Detection of novel prohibited item categories for real-world security inspection,” *Eng. Appl. Artif. Intell.*, vol. 144, p. 110110, 2025, doi: <https://doi.org/10.1016/j.engappai.2025.110110>.
- [42] N. L. Byenkya, N. Rose, and D. Taremwa, “Hybrid YOLOv5s-faster R-CNN detector for complex road-scene object detection,” *Discov. Artif. Intell.*, vol. 6, no. 1, 2026, doi: 10.1007/s44163-026-01253-9.
- [43] E. Ahishakiye, F. Kanobe, D. Taremwa, B. A. Nantongo, L. Nkalubo, and S. Ahimbisibwe, “Enhancing malaria detection and classification using convolutional neural networks-vision transformer architecture,” *Discov. Appl. Sci.*, vol. 7, no. 6, 2025, doi: 10.1007/s42452-025-06704-z.
- [44] P. K. Kisozi, A. Kimwise, M. Kumakech, and M. Balirwa, “Discover Artificial Intelligence Article in Press An explainable and uncertainty aware lesion guided hybrid CNN ViT framework for diabetic retinopathy grading IN IN,” 2026.