Superpowering Open-Vocabulary Object Detectors for X-ray Vision
Supplementary Material

This supplementary material is organized into the following sections: Supp. A outlines ethical considerations related to our work; Supp. B provides the reproducibility statement; Supp. C describes the main characteristics and construction process of our proposed dataset, DET-COMPASS; Supp. D presents additional technical implementation details of RAXO; Supp. E and Supp. F offer further analyses of RAXO's effectiveness; and Supp. G and Supp. H present insights into its performance through qualitative examples.

A. Ethics Statement

We do not anticipate any immediate negative societal impact from our work. However, we encourage future researchers building upon this study to exercise the same level of caution we have maintained, recognizing that RAXO has the potential to be applied for both beneficial and harmful purposes.

The primary motivation behind our research is to enhance open-world perception in X-ray prohibited object detection, addressing the growing diversity of objects in security screening. By improving detection capabilities, our work aims to strengthen public safety in critical security scenarios. Notably, the proposed pipeline and model can be executed entirely on local systems, ensuring that user or institutional privacy remains well protected.

For evaluation, we rely on publicly available, well-established benchmarks, strictly adhering to their licensing terms. Regarding the new DET-COMPASS benchmark introduced in this work, we source images from the publicly available COMPASS-XP [9] X-ray classification dataset, complying fully with its license. Our contribution lies in providing additional bounding box annotations to COMPASS-XP through our human annotation efforts. Importantly, we do not introduce or collect any new images. The human annotation process for DET-COMPASS was conducted following the approval of our institution's ethics board after a thorough committee review.

Lastly, for web-retrieved images, we only retain those explicitly permitted for non-commercial use in this project. Each retrieved image was manually reviewed, ensuring that none contain private information such as human faces or vehicle license plates. We will release our proposed benchmark and prototypes under an appropriate license.

B. Reproducibility Statement

Upon publication, we will make all necessary resources available to facilitate the full reproduction of our experimental results. This includes the source code, precise prompts, and benchmark datasets with their splits. Our proposed framework, RAXO, is developed using *open-source*, *publicly accessible* models and data, reinforcing its reproducibility. A comprehensive breakdown of our pipeline's construction is provided in Sec. 5. Additionally, our supplementary material offers further implementation specifics, including the exact prompts, to assist practitioners in replicating our approach effortlessly. By offering detailed methodological explanations, extensive experimental results, and a fully open-source framework and data, we aim to ensure that our work is easily reproducible, empowering researchers and practitioners to adapt our method across diverse applications.

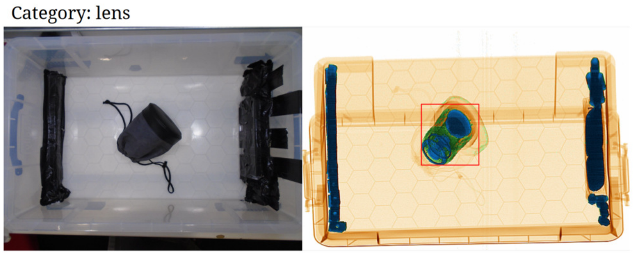

Figure 6. **Occluded RGB object.** In this pair of images, the object `lens` is completely occluded in the RGB image, preventing the annotation of a bounding box

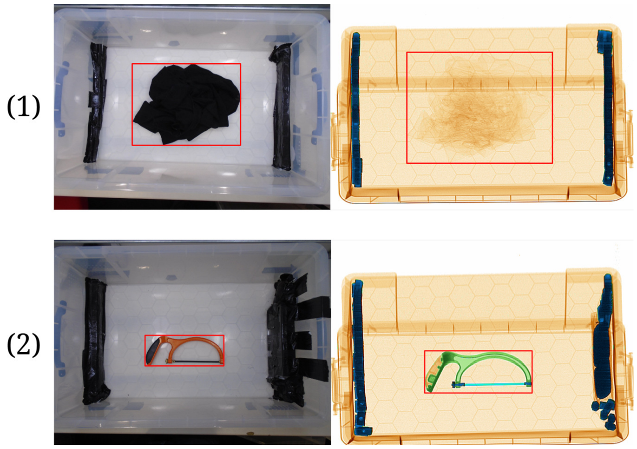

Figure 7. **Visibility attribute.** In (1), the cardigan does not have a discernible signature in the X-ray spectrum, thus `visible=False`. In (2), the hacksaw does, so `visible=True`.

C. DET-COMPASS Details

To construct our new DET-COMPASS dataset, we sourced images from the publicly available COMPASS-XP [9] dataset. Both the images and their metadata are licensed under the Creative Commons Attribution 4.0 International

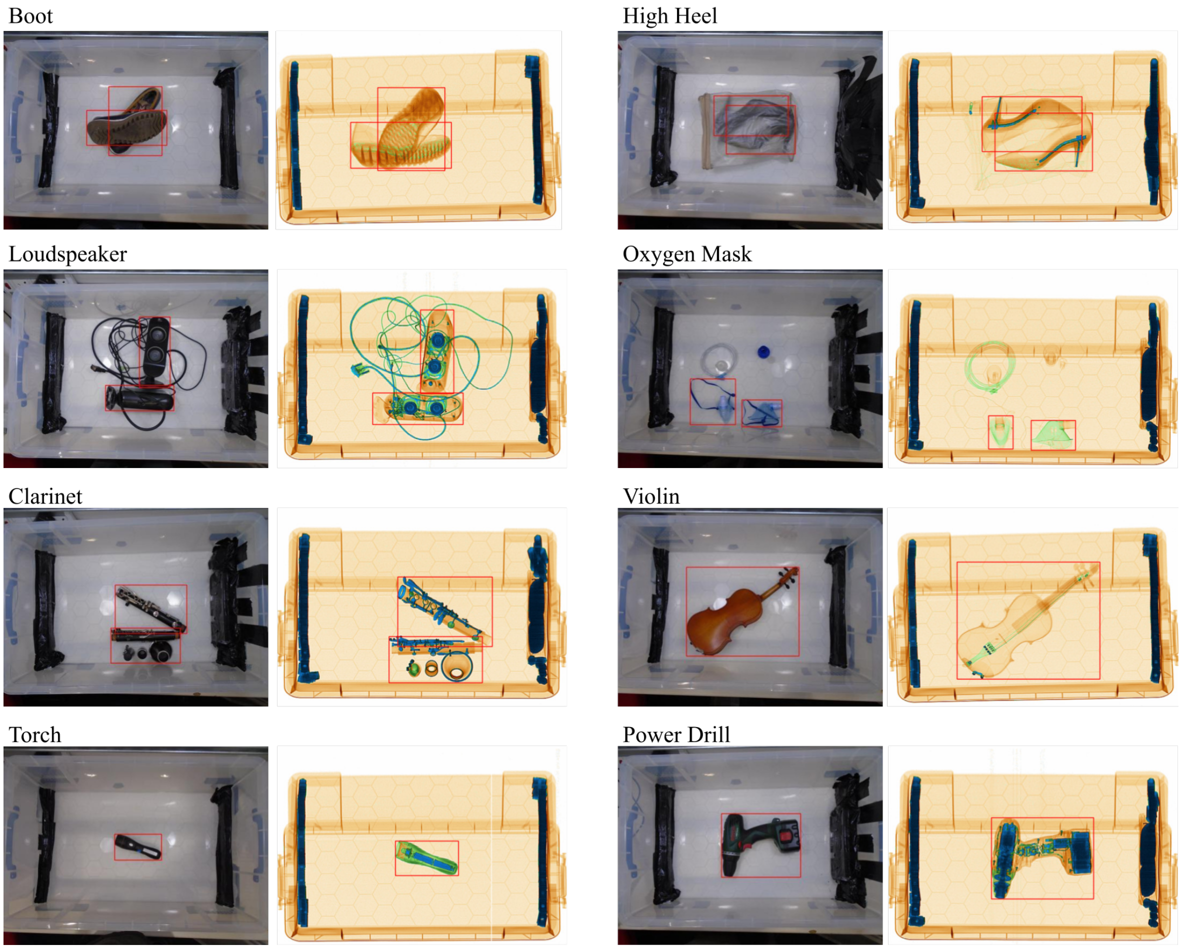

Figure 8. Examples from our **DET-COMPASS dataset**, showing RGB-X-ray pairs with annotated bounding boxes.

License, permitting unrestricted use for research and commercial applications. COMPASS-XP comprises 1,928 image pairs, each consisting of an X-ray image captured with a Gilardoni FEP ME 536 scanner and a corresponding natural image taken with a Sony DSC-W800 digital camera. A key limitation of COMPASS-XP is that it provides only classification labels and the (RGB X-ray) pairs are not spatially aligned.

Our DET-COMPASS dataset builds upon COMPASS-XP by extending the annotations with manually labeled bounding boxes (Fig. 8). The annotation process was conducted by hiring three experts, each responsible for labeling 50% of the RGB-X-ray pairs. To ensure accurate alignment between the RGB and X-ray images, each expert annotated both modalities simultaneously. After completing their respective sets, all three experts reviewed the annotations collectively. One of them acted as a middle ground, overseeing the review process and resolving any remaining discrepancies to ensure annotation consistency.

In total, DET-COMPASS comprises 3,856 annotated images, including 1,928 X-ray and 1,928 RGB images. The average annotation time per image, regardless of modality, was 20 seconds. Given that each expert annotated half of the dataset, the total annotation time amounted to 32.13 hours. The review process required an additional 3 seconds per image, and since all experts participated in reviewing the entire dataset, the total review time was 9.64 hours.

The total number of annotated objects (bounding boxes) in the X-ray images is 1,907, while in the RGB images, it is 1,870. This discrepancy arises because some objects are occluded in the RGB modality, making their localization impossible (Fig. 6). Each annotated object in the X-ray modality includes a *visibility* attribute, indicating whether it produces a discernible signature in the X-ray spectrum. An example of an object marked as visible is shown in Fig. 7(2), while an example of an object marked as non-visible is presented in Fig. 7(1). DET-COMPASS comprises a total of 370 object classes (detailed in Tab. 11), of which 307 contain at least one annotated visible object.

Finally, DET-COMPASS avoids long-tail distribution issues thanks to its uniformly distributed categories, with a low Gini coefficient of $G = 0.26$ (*e.g.*, MS-COCO has $G = 0.57$, where higher G indicates bigger long-tail bias).

D. Further Implementation Details of RAXO

D.1. Pseudo-code of RAXO

In Algorithm 1, we present the pseudocode for the core implementation of RAXO, detailing both the construction of visual descriptors and their use to classify detector proposals.

Algorithm 1: Pseudo-code of RAXO.

```
Input: vocabulary $C^{\text{test}}$; OvOD detector $\mathcal{F}$; test image $\mathbf{I}$;
       in-house database $\mathcal{D}_{\text{XRAY}}^{\text{in-house}}$; web-database $\mathcal{D}_{\text{RGB}}^{\text{web}}$
Output: Detections $\mathcal{T}$ of image $\mathbf{I}$
Initialization: $\mathcal{T} \leftarrow \emptyset$
Initialization: $\mathcal{X} \leftarrow \emptyset$
Initialization: $\mathcal{X}_{bg} \leftarrow \emptyset$
$\mathcal{M} = CreateMaterialDatabase(\mathcal{D}_{\text{XRAY}}^{\text{in-house}})$
/* Visual class descriptors construction */
for class $c \in C^{\text{test}}$ do
    /* VSA refers to the Visual samples
       acquisition pipeline */
    $\mathcal{G}_c^{\text{XRAY}} \leftarrow VSA(c, \mathcal{D}_{\text{XRAY}}^{\text{in-house}})$
    if $\mathcal{G}_c^{XRAY}$ is $\emptyset$ then
        $\widetilde{\mathcal{G}}_c^{\text{web}} \leftarrow VSA(c, \mathcal{D}_{\text{RGB}}^{\text{web}})$
        $\mathcal{G}_c^{\text{web}} = \text{Filter}(\widetilde{\mathcal{G}}_c^{\text{web}}, \mathcal{F}, c, \tau)$
        $\mathcal{A}_m^c = GetMaterialAppareance(\mathcal{M}, c)$
        for sample $\mathbf{u} \in \mathcal{G}_c^{web}$ do
            /* $\Omega$ denotes segmentation */
            $\tilde{\mathbf{u}} = \Omega(\mathbf{u}) \odot (\mathcal{A}_m^c \cdot \mathbf{1})$
            $\mathcal{G}_c^{\text{XRAY}} \leftarrow \mathcal{G}_c^{\text{XRAY}} \cup \{\tilde{\mathbf{u}}\}$
        end
    end
    /* Visual class modeling */
    $\mathcal{X}_c \leftarrow \emptyset$
    for sample $\mathbf{I} \in \mathcal{G}_c^{XRAY}$ do
        $\mathbf{x}_{\mathbf{I}}^{\text{pos}} = Eq.\,(3)$
        $\mathbf{x}_{\mathbf{I}}^{\text{neg}} = Eq.\,(4)$
        $\mathcal{X}_c \leftarrow \mathcal{X}_c \cup \{\mathbf{x}_I^{\text{pos}}\}$
        $\mathcal{X}_{bg} \leftarrow \mathcal{X}_{bg} \cup \{\mathbf{x}_I^{\text{neg}}\}$
    end
    $\mathcal{X}_c \leftarrow \mathcal{X}_c \cup \{Avg(\mathcal{X}_c)\}$
    $\mathcal{X} \leftarrow \mathcal{X} \cup \mathcal{X}_c$
end
/* Detection on image I */
$z = \mathcal{F} \mid \Phi_{RPN}(\mathbf{I})$
$C^{\text{test}'} \leftarrow C^{\text{test}} \cup \{\texttt{background}\}$
for proposal $\mathbf{z_m} \in z$ do
    $\hat{c_m} \leftarrow \arg\max_{c \in C^{\text{test}'}} \max_{\mathcal{X}_c^i \in \mathcal{X}_c} \langle \mathbf{z}_m, \mathcal{X}_c^i \rangle$
    $\hat{\mathbf{b_m}} \leftarrow \mathcal{F} \mid \Phi_{REG}(\mathbf{z_m})$
    /* DCC refers to the Descriptor
       Consistency Criterion */
    if $\hat{c_m}$ is not $\texttt{background}$ and $DCC(\mathbf{z_m}, \mathcal{X})$ then
        $\mathcal{T} \leftarrow \{\hat{c_m} \cup \hat{\mathbf{b_m}}\}$
    end
end
Return: $\mathcal{T}$
```

D.2. Material-Transfer Mechanism

To construct the material database $\mathcal{M}$, we cluster $C^{\text{in-house}}$ into groups of materials identified by a large language model (LLM). The average appearance of objects within each group is used as an estimator of the corresponding material. To perform this clustering, we utilize GPT-4 with the prompt specified in Tab. 12(1).

Once the material database is computed, it can be used to adapt RGB objects to the X-ray modality by inpainting them with their expected material. These expected materials are retrieved from $\mathcal{M}$ using an LLM with the prompt provided in Tab. 12(2).

Material database construction when $\mathcal{D}_{\textbf{XRAY}}^{\textbf{in-house}} = \emptyset$. When no samples are available from $\mathcal{D}_{\text{XRAY}}^{\text{in-house}}$, we construct our material database using the standardized color scheme of security X-ray scans. These scans operate by irradiating objects with X-rays and rendering them in pseudo-colors based on their spectral absorption rates. Typically, three primary pseudo-colors are used [1, 33]: **orange** for organic substances (*e.g.*, food, explosives), **green** for inorganic materials (*e.g.*, laptops, smartphones), and **blue** for metals (*e.g.*, knives, guns). We leverage this modality knowledge to build our material database around these three broad materials.

D.3. Web-retrieval Details

To retrieve images from the web, we utilize the Google Custom Search API [8], configuring specific query parameters to refine the results. We set the search type to images (`searchType: image`) and restrict the results to photos (`imgType: photo`) in common JPEG and PNG formats (`fileType: jpg|png`). To ensure relevance, we limit searches to English-language sources (`lr: lang_en`) and prioritize images from the past seven years (`dateRestrict: y7`).

D.4. In-domain Descriptor Details

In-domain descriptors from $\mathcal{D}_{\text{XRAY}}^{\text{in-house}}$ are built offline by combining the training sets from the six evaluation datasets (PIXray [19], PIDray [34], CLCXray [42], DvXray [20], HiXray [30], and DET-COMPASS) and removing overlapping categories. Combining the datasets ensures a fair evaluation through dataset-agnostic prototypes that capture generic concepts, rather than dataset-specific representations.

D.5. Dataset Colorization

We do not perform color adjustments across datasets, as most do not provide raw density values. However, this does not adversely affect RAXO, since the colorization strategies follow manufacturer-specific yet *consistent palettes* that use similar colors to represent the same materials. These mappings, while differing slightly in hue or intensity, consistently represent the material-specific density and spatial structure necessary for robust detection. Notably, our DET-COMPASS also includes raw density values, enabling more flexible experimentation in future work.

D.6. Complexity Analysis

RAXO is designed to adapt *off-the-shelf* RGB OvOD methods to X-ray without training, making it inherently *modular*. Importantly, most of its components run *offline only once* to build the visual descriptors, requiring roughly 0.7s per class on an NVIDIA A100 GPU. At inference, RAXO simply replaces the text-based classifier of the base OvOD detector with its visual-based classifier, introducing negligible overhead (*e.g.*, 3ms/sample on G-DINO) with complexity $O(n)$ w.r.t. the number of categories.

E. Extended Experimental Results

Maintaining the same experimental setup as in Sec. 6.1, we extend our main results to report AP, AP50, and AP75. Additionally, since the experiments are repeated three times with different random distributions of in-domain and web categories for the intermediate gallery settings, we also report the standard deviation. Tab. 8 show the results. The low standard deviations, combined with RAXO's consistent improvement over all baselines, further validate the effectiveness of RAXO in adapting *off-the-shelf* open-vocabulary detectors to the X-ray modality.

To validate RAXO with an LLM-guided DETR, we also integrated it into LaMI-DETR [4], yielding consistent improvements across all settings (Tab. 9). Finally, to show that the large models in RAXO can be removed or replaced to achieve a desired balance between efficiency and precision, we present an additional ablation in Tab. 10.

F. Per-class AP

Table 6 shows per-class AP on the PIXray dataset for G-DINO. RAXO consistently improves performance, especially on challenging categories with low baseline scores such as *Pressure Vessel* (↑52.3), and *Hammer* (↑54.8). In Tab. 7, we extend the per-category analysis to the DET-COMPASS dataset, analyzing the top-5 classes with the highest and lowest performance gains. RAXO excels on items with distinctive shapes or strong cross-modal color shifts, while struggling with generic-shaped objects that provide limited cues under X-ray.

Category	**G-DINO**	**G-DINO+RAXO**
Pressure Vessel	0.5	**52.8** $_{\uparrow 52.3}$
Bat	70.7	69.7 $_{\downarrow -1.0}$
Gun	31.3	**53.6** $_{\uparrow 22.3}$
Scissors	29.6	**44.3** $_{\uparrow 14.7}$
Razor Blade	0.9	**18.1** $_{\uparrow 17.2}$
Pliers	12.4	**43.5** $_{\uparrow 31.1}$
Dart	0.4	**32.0** $_{\uparrow 31.6}$
Knife	6.2	**10.3** $_{\uparrow 4.1}$
Fireworks	0.0	**2.1** $_{\uparrow 2.1}$
Battery	5.9	**47.7** $_{\uparrow 41.8}$
Saw Blade	3.2	**23.9** $_{\uparrow 20.7}$
Hammer	1.3	**56.1** $_{\uparrow 54.8}$
Screwdriver	1.0	**19.9** $_{\uparrow 18.9}$
Wrench	28.2	**52.3** $_{\uparrow 24.1}$
Lighter	2.0	**26.8** $_{\uparrow 24.8}$
Average	12.9	**36.9** $_{\uparrow +24.0}$

Table 6. **Per-category AP comparison** on the PIXray [19] dataset for G-DINO [18]. RAXO significantly boosts performance across nearly all categories, particularly those with low G-DINO baseline scores.

	Binder	**Milk carton**	**Crayon**	**Hair gel**	**Crowbar**	**Can opener**	**Corkscrew**	**Strainer**	**High heel**	**Compact disc**
G-DINO	0.0	0.1	0.1	0.2	3.0	7.5	1.0	16.7	14.3	1.3
+ **RAXO**	**0.7**$_{\uparrow 0.7}$	**1.3**$_{\uparrow 1.2}$	**1.3**$_{\uparrow 1.2}$	**2.6**$_{\uparrow 2.4}$	**4.2**$_{\uparrow 1.1}$	**88.9**$_{\uparrow 81.4}$	**90.1**$_{\uparrow 89.1}$	**98.2**$_{\uparrow 81.5}$	**98.9**$_{\uparrow 84.6}$	**99.1**$_{\uparrow 97.8}$

Table 7. **Per-category AP on DET-COMPASS** for the top-5 classes with the highest and lowest performance gains. We report AP for G-DINO [18] and G-DINO+RAXO across categories.

G. Qualitative Analysis of the Material Transfer Mechanism

The core challenge that RAXO faces is tackling the domain gap between RGB and X-ray images without training or fine-tuning. The specific component we develop for this purpose is our material-transfer mechanism, whose results compared to a diffusion-based method [7] can be found in Fig. 9.

H. Qualitative Analysis of RAXO

Fig. 10 presents qualitative visualizations of detected X-ray objects before and after applying RAXO with GroundingDINO [18] on the PIXray [19] dataset. For proper visualization, we display detections with a confidence score higher than 0.15 in both cases. These images lead to two key conclusions: (1) RAXO significantly improves the classification of detected proposals. In the baseline images, many objects are correctly localized but misclassified. RAXO successfully corrects these misclassifications by constructing robust visual descriptors. (2) The use of both background descriptors and the Descriptor Consistency Criterion (DCC) effectively eliminates false positives that do not correspond to actual X-ray objects. These observations strongly support the reliability of RAXO.

$\mathcal{G}$	Method	PIXRAY			PIDRAY			CLCXray			COMPASS-XP			HiXray			DVXray		
		AP	AP50	AP75	AP	AP50	AP75	AP	AP50	AP75	AP	AP50	AP75	AP	AP50	AP75	AP	AP50	AP75
	G-DINO [18]	12.9	14.9	13.4	10.9	13.6	11.7	6.7	8.4	7.1	13.4	14.2	13.9	7.0	10.8	8.2	10.0	11.2	10.4
$\mathcal{D}^{\text{in-h}}_{\text{XRAY}}$	100/0	36.9	45.0	39.0	16.5	21.4	17.9	22.2	29.6	24.4	47.9	54.2	48.8	17.1	27.2	19.4	22.6	26.6	24.1
	80/20	33.8±0.6	40.9±0.9	35.5±0.6	15.4±0.4	19.8±0.6	16.6±0.4	18.0±2.1	23.7±2.3	19.5±2.2	41.0±2.2	46.2±2.4	41.7±2.2	14.5±0.6	23.5±1.0	16.3±0.6	21.0±0.6	24.8±0.9	22.3±0.6
↓	50/50 + **RAXO**	25.4±2.0	31.2±1.9	26.7±2.0	15.5±0.9	19.8±1.0	16.8±1.0	17.0±1.8	22.9±3.2	18.7±2.3	31.4±0.7	35.3±0.9	32.1±0.6	13.4±0.1	21.3±0.1	15.3±0.2	16.1±1.8	18.8±2.3	17.0±2.0
	20/80	21.6±0.6	26.1±1.1	22.6±0.6	13.9±0.5	17.9±0.7	14.9±0.6	10.0±0.4	13.1±1.5	10.7±0.7	20.5±0.6	22.9±0.7	21.1±0.7	9.8±1.0	15.8±1.4	11.1±1.2	15.0±1.0	17.2±1.1	15.8±1.2
$\mathcal{D}^{\text{web}}_{\text{RGB}}$	0/100	16.1	19.8	16.8	13.4	17.1	14.3	7.1	9.7	7.5	14.0	15.4	14.5	7.9	13.0	8.7	12.4	14.1	12.9
	Detic [44]	9.3	11.6	9.5	7.1	9.7	7.6	4.7	7.3	4.6	11.5	13.4	13.3	4.8	8.6	5.2	7.0	8.5	7.5
$\mathcal{D}^{\text{in-h}}_{\text{XRAY}}$	100/0	27.3	34.5	28.2	11.3	15.8	12.2	14.0	20.6	14.7	35.3	39.9	35.4	14.2	23.9	15.5	19.4	23.9	21.2
	80/20	23.9±1.3	30.2±1.5	24.6±1.3	10.8±0.1	15.0±0.2	11.7±0.1	12.3±1.6	18.1±1.8	12.8±1.9	30.7±1.4	34.4±1.3	30.8±1.5	12.1±1.1	20.8±1.8	13.1±1.2	18.0±2.2	22.1±2.6	19.7±2.4
↓	50/50 + **RAXO**	19.5±1.6	24.8±1.9	20.1±1.7	10.3±0.3	14.3±0.3	11.0±0.3	9.2±1.2	13.5±2.3	9.5±1.2	24.4±2.7	27.1±2.7	24.8±2.6	11.0±0.9	18.9±1.3	11.9±1.2	14.6±1.1	17.9±1.2	15.9±1.2
	20/80	15.2±0.9	19.4±0.9	15.5±1.0	9.6±0.1	13.3±0.2	10.3±0.2	8.0±0.1	12.5±0.1	8.0	16.4±1.0	18.3±1.0	16.4±1.0	9.9±0.8	16.8±1.4	10.7±0.9	12.7±0.6	15.5±0.8	13.9±0.7
$\mathcal{D}^{\text{web}}_{\text{RGB}}$	0/100	13.4	16.8	13.6	9.1	12.6	9.8	5.2	8.1	5.1	11.9	13.1	12.1	7.9	13.8	8.4	9.4	11.4	10.1
	CoDet [21]	7.3	8.7	7.6	5.7	7.6	6.2	3.1	5.7	2.7	8.4	8.9	8.7	3.4	5.9	3.7	5.6	6.8	6.0
$\mathcal{D}^{\text{in-h}}_{\text{XRAY}}$	100/0	27.9	33.6	29.2	10.3	14.6	10.9	14.8	22.4	15.9	35.8	41.0	36.7	13.2	22.0	14.8	17.6	21.7	19.0
	80/20	25.1±1.5	30.2±1.7	26.2±1.7	9.5±0.3	13.4±0.5	10.1±0.3	12.0±1.9	18.3±2.8	12.7±2.1	32.2±0.9	36.5±1.5	33.1±0.6	11.7±1.3	19.4±2.2	13.2±1.5	15.4±1.4	18.8±1.7	16.7±1.6
↓	50/50 + **RAXO**	20.0±0.7	24.1±0.9	20.8±0.7	9.5±0.5	13.4±0.7	10.1±0.5	9.2±1.4	14.2±2.1	9.6±1.7	24.0±0.2	26.7±0.3	24.7±0.2	9.9±0.4	16.7±0.8	11.1±0.4	11.5±0.8	14.2±1.1	12.4±0.8
	20/80	14.8±2.4	17.8±2.8	15.3±2.5	8.5±0.3	11.9±0.4	9.0±0.4	5.1±1.4	9.0±2.5	5.0±1.6	17.8±0.7	19.4±0.9	18.2±0.6	8.1±0.6	13.8±1.0	8.8±0.6	9.4±1.5	11.3±1.8	10.1±1.6
$\mathcal{D}^{\text{web}}_{\text{RGB}}$	0/100	11.5	14.0	11.9	8.1	11.3	8.7	4.0	7.1	3.8	12.2	13.0	12.6	6.5	11.2	7.1	6.9	8.3	7.5
	VLDet [15]	9.8	12.1	10.3	6.9	9.4	7.4	4.4	7.8	4.0	10.6	11.4	10.8	5.1	9.0	5.5	7.4	9.2	8.1
$\mathcal{D}^{\text{in-h}}_{\text{XRAY}}$	100/0	32.3	40.1	34.0	11.7	16.8	12.6	15.4	23.3	15.9	36.4	41.4	37.2	14.8	24.5	16.3	20.1	25.1	22.0
	80/20	29.2±1.2	36.3±1.2	30.7±1.3	11.0±0.3	15.7±0.3	11.7±0.3	12.7±0.5	19.6±1.2	13.0±0.5	31.8±0.8	36.0±1.0	32.5±0.9	13.1±1.2	21.8±1.9	14.3±1.3	16.8±0.2	21.0±0.1	18.4±0.1
↓	50/50 + **RAXO**	24.0±1.5	29.9±1.7	25.2±1.5	10.4±0.7	14.6±1.0	11.1±0.8	11.1±1.1	16.9±0.4	11.5±1.7	23.7±0.9	26.5±0.8	24.3±1.1	11.2±1.5	19.0±2.1	12.1±1.9	12.1±0.5	15.0±0.4	13.2±0.4
	20/80	21.6±1.0	26.8±0.9	22.6±1.0	9.4±0.3	13.3±0.4	10.1±0.3	5.2±0.1	9.1±0.2	4.8±0.0	16.2±0.9	18.2±1.2	16.6±1.0	9.3±0.2	15.9±0.2	9.9±0.3	10.6±0.5	13.1±0.6	11.5±0.5
$\mathcal{D}^{\text{web}}_{\text{RGB}}$	0/100	14.1	17.8	14.5	8.9	12.5	9.5	4.4	8.1	3.9	11.1	12.2	11.4	8.3	14.5	8.7	9.0	11.0	9.8

Table 8. **X-ray OvOD performance under the Cross-Modality Transfer Evaluation (CMTE) setting** on DET-COMPASS (ours), PIXray [19], PIDray [34], CLCXray [42], DvXray [20], and HiXray [30] datasets. We integrate RAXO into different baselines using different gallery $\mathcal{G}$ compositions, from using only $\mathcal{D}^{\text{in-house}}_{\text{XRAY}}$ data (100/0) to exclusively $\mathcal{D}^{web}_{\text{RGB}}$ samples (0/100). RAXO consistently improves the performance of all baseline OvOD detectors across every dataset. We report the AP, AP50 and AP75. We also include the deviations because each experiment is repeated three times with different random distributions of in-domain and web categories for the intermediate gallery settings.

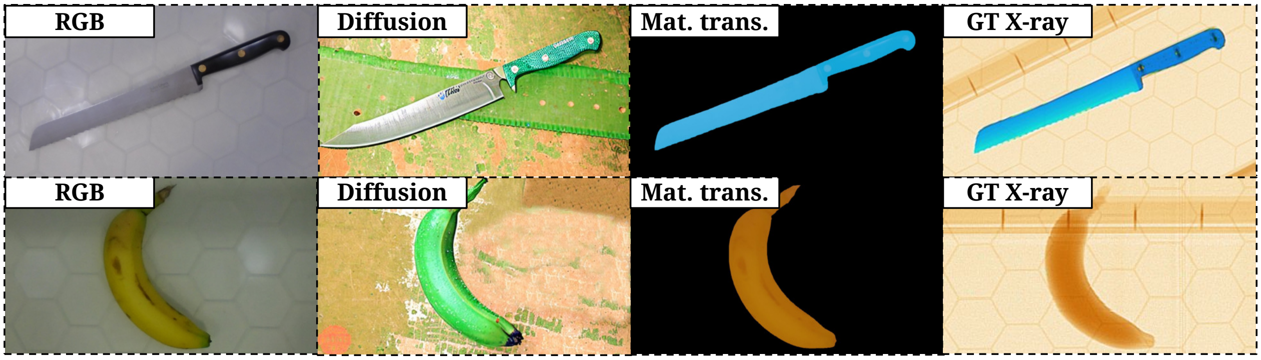

Figure 9. **Qualitative comparison** between our material-transfer mechanism and a diffusion-based method [7].

$\mathcal{G}$	Method	D-COMP.	PIXray	PIDray	CLCXray	DvXray	HiXray
	LaMI-DETR	11.3	13.6	8.0	4.0	9.7	6.3
100/0		**31.9**$_{\uparrow 20.6}$	**25.7**$_{\uparrow 12.1}$	**13.1**$_{\uparrow 5.1}$	**18.7**$_{\uparrow 14.7}$	**18.1**$_{\uparrow 8.4}$	**9.7**$_{\uparrow 3.4}$
80/20		**27.2**$_{\uparrow 15.9}$	**23.2**$_{\uparrow 9.6}$	**12.0**$_{\uparrow 4.0}$	**16.2**$_{\uparrow 12.2}$	**16.0**$_{\uparrow 6.3}$	**8.3**$_{\uparrow 2.0}$
50/50	+ **RAXO**	**22.0**$_{\uparrow 10.7}$	**15.9**$_{\uparrow 2.3}$	**12.1**$_{\uparrow 4.1}$	**15.1**$_{\uparrow 11.1}$	**12.8**$_{\uparrow 3.1}$	**7.6**$_{\uparrow 1.3}$
20/80		**15.7**$_{\uparrow 4.4}$	**15.2**$_{\uparrow 1.6}$	**10.8**$_{\uparrow 2.8}$	**6.8**$_{\uparrow 2.8}$	**11.9**$_{\uparrow 2.2}$	**7.0**$_{\uparrow 0.7}$
0/100		**11.5**$_{\uparrow 0.2}$	**14.8**$_{\uparrow 1.2}$	**10.8**$_{\uparrow 2.8}$	**6.2**$_{\uparrow 2.2}$	**10.5**$_{\uparrow 0.8}$	**6.5**$_{\uparrow 0.2}$

Table 9. **X-ray OvOD performance under the Cross-Modality Transfer Evaluation (CMTE) setting** on DET-COMPASS (ours), PIXray [19], PIDray [34], CLCXray [42], DvXray [20], and HiXray [30] datasets. We integrate RAXO into LaMI-DETR [4] using different gallery $\mathcal{G}$ compositions, from using only $\mathcal{D}^{\text{in-house}}_{\text{XRAY}}$ data (100/0) to exclusively $\mathcal{D}^{web}_{\text{RGB}}$ samples (0/100). RAXO consistently improves the performance of LaMI-DETR.

	Segment.	LLM	Features	PIXray (*50/50*)		
				AP	AP50	AP75
G-DINO [18]				12.9	14.9	13.4
+ **RAXO**	SAM 2	GPT-4	DINOv2	**25.4**$_{\uparrow 12.5}$	**31.2**$_{\uparrow 16.3}$	**26.7**$_{\uparrow 13.3}$
	–	GPT-4	DINOv2	**22.0**$_{\uparrow 9.1}$	**27.3**$_{\uparrow 12.4}$	**22.7**$_{\uparrow 9.3}$
	SAM 2	–	DINOv2	**20.8**$_{\uparrow 7.9}$	**24.1**$_{\uparrow 9.2}$	**21.4**$_{\uparrow 8.0}$
	SAM 2	GPT-4	DINO	**22.2**$_{\uparrow 9.3}$	**27.6**$_{\uparrow 12.7}$	**22.9**$_{\uparrow 9.5}$
	SAM 2	LLaMA-3	DINOv2	**24.7**$_{\uparrow 11.8}$	**30.1**$_{\uparrow 15.2}$	**26.1**$_{\uparrow 12.7}$
	SAM	GPT-4	DINOv2	**25.1**$_{\uparrow 12.2}$	**31.0**$_{\uparrow 16.1}$	**26.4**$_{\uparrow 13.0}$

Table 10. **Ablation study of RAXO components on the PIXray [19] dataset (50/50 setting).** We integrate RAXO into G-DINO and analyze the impact of segmentation models, language models, and visual features. Results show that each component incrementally boosts performance, with the full RAXO configuration yielding the best results.

DET-COMPASS Categories						
abacus	abaya	amplifier	analog watch	apron	baby monitor	backpack
bag of sweets	baking dish	ballpoint	banana	Band Aid	baseball bat	baseball cap
bath towel	bathing cap	beanie	beer bottle	beer glass	bell pepper	belt
bib	bicycle helmet	bikini	binder	binoculars	bird feeder	biscuits
blowtorch	boardgame	book	book jacket	boot	bow tie	bowl
bowler hat	box cutter	bracelet	brassiere	bread knife	brush	bumbag
butternut squash	cable	caliper	camcorder	camera	can opener	candle
canned food	capo	cardigan	cards	carving knife	cassette	cassette player
cd drive	cellular telephone	cereal	chain	charger	chewing gum	chisel
chocolate	chocolate sauce	Christmas stocking	cigarettes	clarinet	coat hanger	cocktail shaker
coffee mug	coffeepot	colander	comb	combination lock	comic book	compact disc
condoms	corkscrew	cotton buds	cotton wool	cowboy hat	craft knife	crayon
crisps	crossword puzzle	crowbar	cucumber	dagger	denture	deodorant
diaper	digital watch	dinner jacket	dishrag	dressing gown	dvd player	e cigarette
e liquid	electric fan	electric toothbrush	empty	envelope	espresso maker	extension cord
face powder	fascinator	feather boa	first aid kit	floss	flute	fork
French loaf	frisbee	frying pan	fur coat	gaffer tape	game console	gameboy
gas canister	glove	glue gun	goggles	hacksaw	hair clippers	hair gel
hair spray	hairbrush	hammer	hand blower	handkerchief	hard disc	harmonica
hatchet	headphones	hearing aid	high heel	hook	hourglass	ipad
iPod	iron	jean	jersey	jewellery box	jigsaw puzzle	joystick
jumper	kettle	keys	kimono	kindle	kiwi	knee pad
knife	lab coat	ladle	lampshade	laptop	laser pointer	leather jacket
lemon	lens	lens cap	letter opener	lighter	lime	lipstick
lotion	loudspeaker	loupe	magazine	magnetic compass	maillot	mallet
marker	mask	matchstick	measuring cup	microphone	milk can	milk carton
mitten	mixing bowl	modem	mortar	mosquito net	mouse	mousetrap
mouthwash	multimeter	music stand	nail	nail clippers	nail file	nail scissors
necklace	notebook	orange	oxygen mask	padlock	paint can	paintbrush
pajama	paper towel	passport	pasta	pencil	pencil box	pencil sharpener
penknife	pepper grinder	perfume	pick	pickaxe	piggy bank	pill bottle
pillow	plane	plastic bag	plate	plate rack	pliers	plunger
Polaroid camera	polo shirt	pomegranate	poncho	pop bottle	pot	power drill
power socket	power supply	prayer rug	quill	quilt	quilted jacket	radio
rasp	razor	razor blades	recorder	red wine	reflex camera	remote control
rice	roll of sweets	roller skate	rubber eraser	rubber gloves	rubik cube	rugby ball
rugby shirt	rule	running shoe	safety pin	salad bowl	saltshaker	sandal
sandwich	sarong	saucepan	saw	sax	scale	scarf
scissors	screw	screwdriver	secateurs	sellotape	sewing machine	shampoo
shaver	shawl	shirt	shorts	shovel	shower cap	sieve
ski mask	skipping rope	sleeping bag	slide	slotted spoon	smartphone	snorkel
soap	soap dispenser	sock	solder	soldering iron	sombrero	soup bowl
spatula	spectacles	spirit level	splitter block	spoon	spotlight	staple gun
stapler	stethoscope	stockings	stole	stopwatch	strainer	strings
stylophone	suit	sunglasses	sunscreen	swab	sweatshirt	swimming trunks
switch	syringe	table lamp	tampon	tape measure	tea towel	teapot
teaspoon	teddy	telephone	telescope	tennis ball	thermals	thermometer
thermos	tin of sweets	toaster	toilet tissue	toner cartridge	toothbrush	toothpaste
top hat	torch	tracksuit	tray	tripod	tuner	ukulele
umbrella	underpants	vacuum	vase	velvet	vinyl record	violin
waffle iron	walking boot	wall clock	wallet	washbag	water bottle	water jug
wellington boot	wet wipes	whetstone	whistle	wig	wineglass	wire wool
wirecutter	wok	wooden spoon	wool	wrench	wrist guard	

Table 11. Category names of DET-COMPASS.

(1): Material-database clustering prompt

“You are a computer expert specializing in material classification. Your task is to analyze a given list of objects, determine their primary material composition, and group them accordingly.

Instructions:
Identify the main materials present among the objects (e.g., metal, organic, inorganic, plastic, ceramic, etc.). Assign each object to the most appropriate material category. Each object should belong to only one category based on its primary composition. Return the results in JSON format, where the keys are material categories, and the values are lists of objects belonging to those categories.

Example:
Input: Objects: gun, bat, pressure vessel, beer glass, fur coat, lemon
Expected Output (JSON):
metal: [gun, bat],
inorganic: [pressure vessel, beer glass],
organic: [fur coat, lemon]

Now, classify the following list of objects: $\{D^{\text{in-house}}\}$. Return only the json format.”

(2): Object material identification prompt

“You are a computer vision assistant. Given a $\{object\}$, classify it into one of the following materials: $\{\mathcal{M}.materials\_names\}$. Return only the material. You must always select one.”

Table 12. **Prompts used for material clustering and retrieval**. (1) The clustering prompt provided to GPT-4 to group $C^{\text{in-house}}$ into material categories. (2) The retrieval prompt used to query $\mathcal{M}$ and infer the expected material of unknown RGB objects.

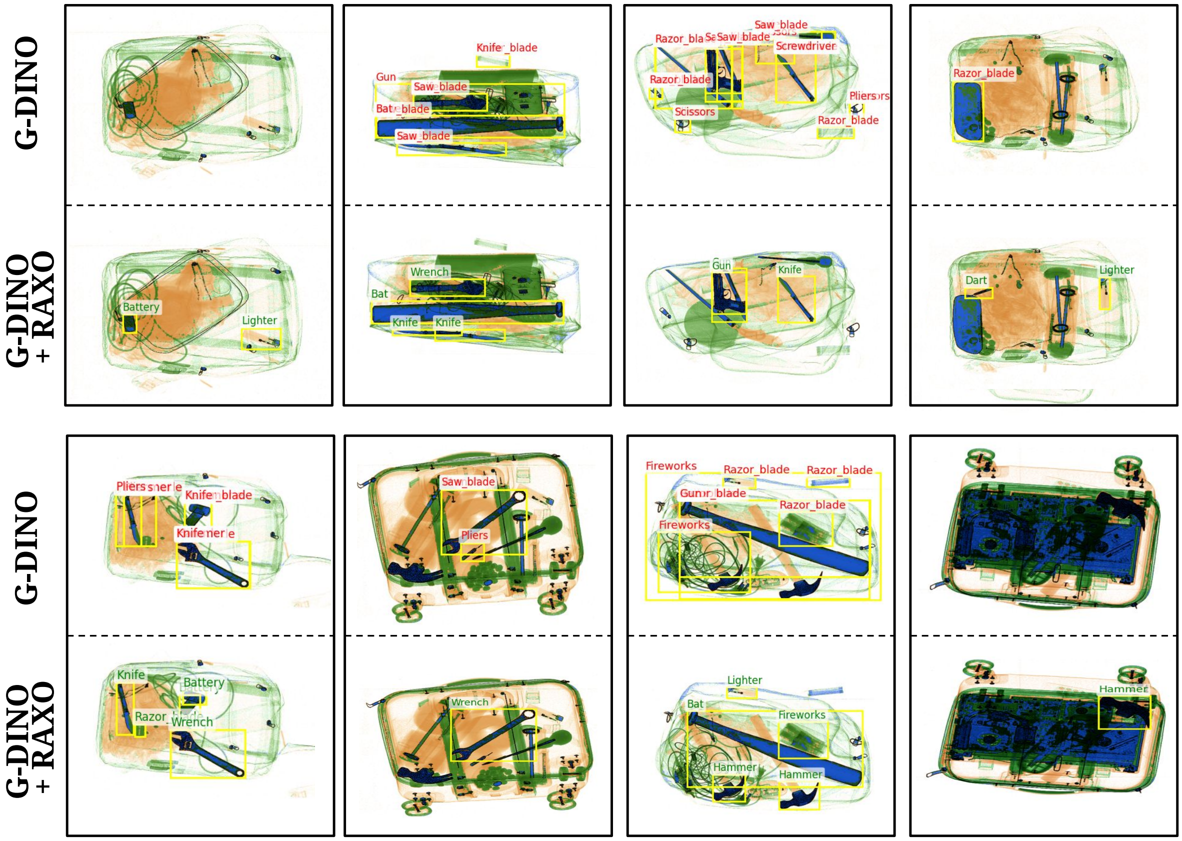

Figure 10. **Qualitative comparison** of G-DINO [18] and G-DINO+RAXO.

References

[1] Samet Akcay and Toby Breckon. Towards automatic threat detection: A survey of advances of deep learning within x-ray security imaging. *Pattern Recognition*, 122:108245, 2022. 2, 3

[2] An Chang, Yu Zhang, Shunli Zhang, Leisheng Zhong, and Li Zhang. Detecting prohibited objects with physical size constraint from cluttered x-ray baggage images. *Knowledge-Based Systems*, 2022. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 6

[4] Penghui Du, Yu Wang, Yifan Sun, Luting Wang, Yue Liao, Gang Zhang, Errui Ding, Yan Wang, Jingdong Wang, and Si Liu. Lami-detr: Open-vocabulary detection with language model instruction. In *ECCV*, 2024. 4, 5

[5] Ruohuan Fang, Guansong Pang, and Xiao Bai. Simple image-level classification improves open-vocabulary object detection. In *AAAI*, 2024. 2

[6] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *ECCV*, 2022. 2

[7] Junyao Gao, Yanchen Liu, Yanan Sun, Yinhao Tang, Yanhong Zeng, Kai Chen, and Cairong Zhao. Styleshot: A snapshot on any style. *arXiv:2407.01414*, 2024. 4, 7, 8, 5

[8] Google LLC. Custom search JSON API reference. `https://developers.google.com/custom-search/v1`, 2024. Accessed: 2025-03-07. 6, 3

[9] Lewis D. Griffin, Matthew Caldwell, and Jerone T. A. Andrews. COMPASS-XP. Zenodo, 2019. 3, 1

[10] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for open-vocabulary segmentation. In *ECCV*, 2024. 5

[11] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv:2209.15639*, 2022. 2

[12] Liangqi Li, Jiaxu Miao, Dahu Shi, Wenming Tan, Ye Ren, Yi Yang, and Shiliang Pu. Distilling detr with visual-linguistic knowledge for open-vocabulary object detection. In *ICCV*, 2023. 2

[13] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, 2022. 2, 7

[14] Mingyuan Li, Tong Jia, Hao Wang, Bowen Ma, Hui Lu, Shuyang Lin, Da Cai, and Dongyue Chen. Ao-detr: Anti-overlapping detr for x-ray prohibited items detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 1

[15] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. In *ICLR*, 2023. 1, 2, 6, 7, 5

[16] Shuyang Lin, Tong Jia, Hao Wang, Bowen Ma, Mingyuan Li, and Dongyue Chen. Detection of novel prohibited item categories for real-world security inspection. *Eng. Appl. Artif. Intell.*, 2025. 1, 2, 7

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 6

[18] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In *ECCV*, 2024. 1, 2, 6, 7, 8, 4, 5

[19] Bowen Ma, Tong Jia, Min Su, Xiaodong Jia, Dongyue Chen, and Yichun Zhang. Automated segmentation of prohibited items in x-ray baggage images using dense de-overlap attention snake. *TMM*, 2022. 2, 3, 6, 7, 8, 4, 5

[20] Bowen Ma, Tong Jia, Mingyuan Li, Songsheng Wu, Hao Wang, and Dongyue Chen. Towards dual-view x-ray baggage inspection: A large-scale benchmark and adaptive hierarchical cross refinement for prohibited item discovery. *IEEE TIFS*, 2024. 1, 3, 6, 7, 5

[21] Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. CoDet: Co-occurrence guided region-word alignment for open-vocabulary object detection. In *NeurIPS*, 2023. 2, 6, 7, 5

[22] Caijing Miao, Lingxi Xie, Fang Wan, Chi Su, Hongye Liu, Jianbin Jiao, and Qixiang Ye. Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images. In *CVPR*, 2019. 2

[23] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *ECCV*, 2022. 2

[24] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 6

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2023. 6

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[27] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv:2408.00714*, 2024. 6

[28] Haifeng Sima, Bailiang Chen, Chaosheng Tang, Yudong Zhang, and Junding Sun. Multi-scale feature attention-detection transformer: Multi-scale feature attention for security check object detection. *IET Computer Vision*, 2024. 2

[29] Archana Singh and Dhiraj. Advancements in machine learning techniques for threat item detection in x-ray images: a comprehensive survey. *Int. J. Multim. Inf. Retr.*, 2024. 1

[30] Renshuai Tao, Yanlu Wei, Xiangjian Jiang, Hainan Li, Haotong Qin, Jiakai Wang, Yuqing Ma, Libo Zhang, and Xianglong Liu. Towards real-world x-ray security inspection: A high-quality benchmark and lateral inhibition module for prohibited items detection. In *ICCV*, 2021. 2, 3, 6, 7, 5

[31] Renshuai Tao, Hainan Li, Tianbo Wang, Yanlu Wei, Yifu Ding, Bowei Jin, Hongping Zhi, Xianglong Liu, and Aishan Liu. Exploring endogenous shift for cross-domain detection: A large-scale benchmark and perturbation suppression network. In *CVPR*, 2022. 2, 3

[32] Renshuai Tao, Tianbo Wang, Ziyang Wu, Cong Liu, Aishan Liu, and Xianglong Liu. Few-shot x-ray prohibited item detection: A benchmark and weak-feature enhancement network. In *ACMMM*, 2022. 1, 3

[33] Divya Velayudhan, Taimur Hassan, Ernesto Damiani, and Naoufel Werghi. Recent advances in baggage threat detection: A comprehensive and systematic survey. *ACM Computing Surveys*, 55(8):1–38, 2022. 2, 3

[34] Boying Wang, Libo Zhang, Longyin Wen, Xianglong Liu, and Yanjun Wu. Towards real-world prohibited item detection: A large-scale x-ray benchmark. In *ICCV*, 2021. 2, 3, 6, 7, 5

[35] Ruxue Wang, Yuliang Shi, and Mingyu Cai. Optimization and research of suspicious object detection algorithm in x-ray image. In *2023 IEEE 6th Information Technology,Networking,Electronic and Automation Control Conference (ITNEC)*, 2023. 2

[36] Yanlu Wei, Renshuai Tao, Zhangjie Wu, Yuqing Ma, Libo Zhang, and Xianglong Liu. Occluded prohibited items detection: An x-ray security inspection benchmark and deocclusion attention module. In *ACMMM*, 2020. 2

[37] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *CVPR*, 2023. 2, 7

[38] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. In *ICLR*, 2024. 2

[39] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *CVPR*, 2023. 2

[40] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pretraining for open-world detection. In *NeurIPS*, 2022. 2

[41] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *CVPR*, 2023. 2

[42] Cairong Zhao, Liang Zhu, Shuguang Dou, Weihong Deng, and Liang Wang. Detecting overlapped objects in x-ray security imagery by a label-aware mechanism. *IEEE TIFS*, 2022. 2, 3, 6, 7, 5

[43] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, 2022. 2

[44] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 1, 2, 3, 6, 7, 5

[45] Chaoyang Zhu and Long Chen. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE TPAMI*, 2024. 3, 6