Falcon : Functional Assembly and Language for Compositional Reasoning in X-ray

Yonathan Michael[†⋆], Mohamad Alansari[†⋆], Natnael Takele[†], Andreas Henschel[†‡], and Naoufel Werghi[†‡]

[†]Computer Science Department [‡]Center of Cyber-physical Systems (C2PS)
Khalifa University
{100053679,100061914,100058082,andreas.henschel,naoufel.werghi}@ku.ac.ae
https://yonathan-kiflom.github.io/FALCON/

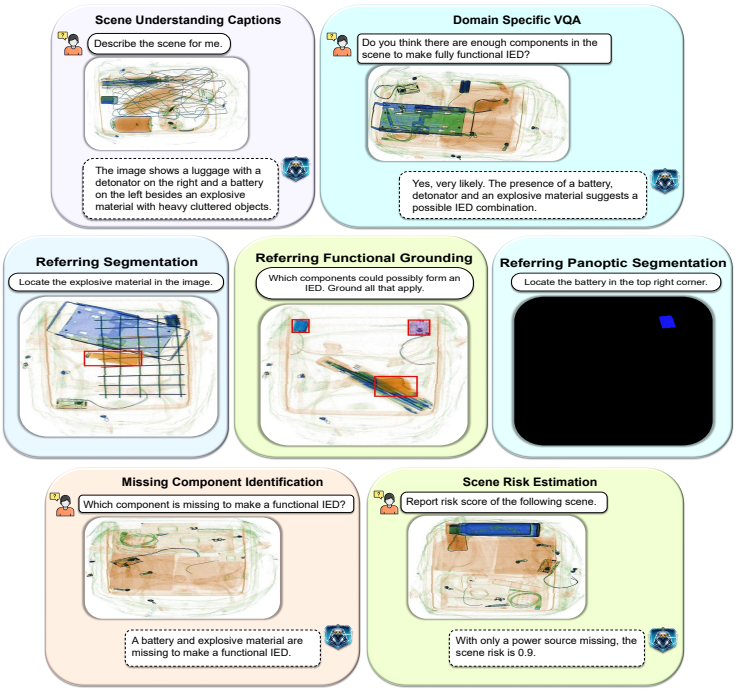

Fig. 1: Falcon enables segmentation-aware functional threat reasoning in X-ray imagery. It performs instance grounding, component presence recognition, referring functional grounding and more, while generating grounded and domain-aware natural-language explanations. Falcon reasons over spatially distant components of IED under heavy clutter, supporting structured threat analysis beyond object detection.

Abstract. Conventional vision-language models are largely object-centric, focusing on detecting and describing individual entities. In safety-critical X-ray baggage screening, however, threat often emerges not from a single object but from the functional compatibility of spatially dispersed components, such as batteries, detonators, and explosive charges. We formalize this setting as *compositional threat reasoning*, where risk is modeled

⋆ Equal contribution.

as a relational property of grounded regions rather than an independent detection outcome. We introduce **Falcon**, a multimodal framework that abstracts segmentation-aware region features into a structured safety state capturing component presence, pairwise functional compatibility, and scene-level risk. This structured representation is injected into the language model as an explicit intermediate interface, encouraging relationally consistent and safety-aware reasoning. To evaluate this problem, we present **Falcon-X**, a benchmark that unifies dense grounding with structured supervision over component completeness and risk inference in cluttered X-ray imagery. Experiments show that while existing multimodal models adapt to appearance, they struggle with compositional safety reasoning. Falcon improves functional grounding and produces more coherent threat assessments, establishing compositional safety reasoning as a distinct evaluation paradigm for multimodal systems.

1 Introduction

X-ray baggage screening is a safety-critical task characterized by heavy clutter, material transparency, and severe object overlap [33]. Recent deep learning approaches have significantly advanced prohibited-item detection [2, 13, 30] on large-scale X-ray benchmarks [21, 22, 31, 36, 38], and multimodal models have enabled captioning and VQA in this domain [32]. However, existing systems remain fundamentally object-centric and threat is inferred from the presence of individual categories. In operational settings, however, risk is often *compositional*. Improvised explosive devices (IEDs) may be transported in dismantled form, with batteries, detonators, and explosive charges spatially separated within a bag. Individually benign, these components become hazardous only through *functional compatibility*. The central question is therefore not object recognition, but relational safety inference:

Can spatially separated components under severe superposition collectively constitute a functional threat?

Addressing this setting requires structured reasoning over multiple grounded regions and their interactions. Current X-ray benchmarks evaluate object-level detection, while multimodal large language models (MLLMs) [3, 11, 12, 16, 19, 25, 27, 37] rely on implicit attention without explicit supervision over inter-component compatibility. Consequently, relationally consistent safety inference remains largely unexplored in cluttered X-ray imagery.

We formalize *compositional threat reasoning* as a structured multimodal prediction problem in which safety is modeled as a relational property of grounded components. We introduce **Falcon**, a segmentation-aware multimodal framework that maps region-level features to an explicit safety state comprising component presence, pairwise functional compatibility, and calibrated scene-level risk. This structured state is injected into the language model via a **Structured Safety Adapter (SSA)**, enforcing relational consistency between perception and reasoning (Fig. 1).

Table 1: Comparison of Falcon-X with existing X-ray security benchmarks. DC denotes explicit modeling of dismantled components. Falcon-X is the only dataset that jointly supports dense grounding, multimodal understanding, and structured functional threat reasoning.

Dataset	# Samples	DC	Annotations				Task Suite								Avail.
			Labels	Bbox	Mask	Caption	RefSeg	RefPan	Pan	VQA	RL	FG	GG	SA	
GDXray (JNDE'15) [17]	8,150	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
SIXray (CVPR'19) [22]	8,929	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
OPIXray (ACMMM'20) [34]	46,642	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
HiXray (ICCV'21) [31]	45,364	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
CLCxray (ITIFS'22) [38]	9,565	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
PIXray (ITMM'22) [18]	5,046	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
EDS (CVPR'22) [29]	14,219	✗	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓
LPIXray (CIPAE'23) [7]	60,950	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
PIDray (IJCV'23) [36]	47, 677	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
DvXray (ITIFS'24) [17]	5,496	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
STCray (CVPR'25) [32]	46,642	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓
Falcon-X	**6,911**	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

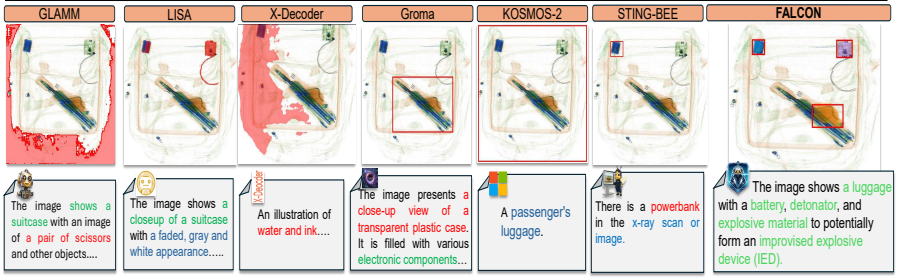

Fig. 2: Qualitative comparison of structured threat reasoning. Existing models, including X-ray domain VLMs, do not model functional relation of components. RED: indicates hallucinations, BLUE: indicates weak semantic relevance, and GREEN: indicates correct description.

To evaluate this setting, we introduce **Falcon-X**, a benchmark of ∼7,000 real X-ray scans with instance-level annotations for key IED components. Beyond object detection, Falcon-X provides structured supervision for component completeness and calibrated risk, enabling evaluation of relational grounding and compositional safety reasoning under severe superposition. We benchmark Falcon-X across general-purpose and domain-adapted VLMs (Fig. 2).
Our contributions are threefold:

- We formalize compositional threat reasoning in X-ray imagery, modeling safety as a relational property of spatially grounded components.
- We introduce Falcon-X, the first benchmark coupling dense grounding with structured supervision for functional completeness and safety-consistent risk inference.
- We propose Falcon, a structured multimodal framework that injects an explicit safety state into region-level MLLMs to improve relationally coherent threat assessment.

2 Related Work

General MLLMs: Multimodal large language models (MLLMs) extend language models to vision by coupling visual encoders with language backbones.

Early methods [12, 26] demonstrate the effectiveness of large-scale image-text pretraining for transferrable multimodal representations. Instruction tuned extensions [16,39] build conversational interfaces on top of frozen pretrained vision encoders to answer complex queries about whole images. To enable explicit target localization, recent works [3, 19, 25] incorporate region-level representations by encoding bounding boxes as tokens and [37] achieves that by extracting region features via pooling. [11, 27] further extend language models to generate pixel-level masks for fine-grained localization. Despite these advances, such MLLMs have been evaluated mainly on everyday images and visually separable scenarios, and they do not target safety-critical tasks to explicitly reason over functional relationships or to detect disassembled threat components under heavy occlusion.

X-ray baggage analysis: X-ray security images pose unique challenges due to material transparency, object overlap and heavy clutter. Large-scale benchmarks [21, 22, 31, 34, 36] have enabled progress in detecting prohibited items under the above conditions. These datasets, predominantly follow a closed-set, object-centric paradigm, focusing on intact common threat categories such as guns, knives and scissors. Recent MLLM work introduced image-text paired dataset STCray [32] for extended tasks of image captioning, target localization and VQA in X-ray domain. However, existing benchmarks [23] evaluate threats at the object level and do not explicitly model spatially scattered configurations of prohibited items in which risk arises from the functional composition of distributed components. Missing-part reasoning, functional sufficiency assessment and structured safety evaluation remain largely unexplored in this domain.

3 Problem Setup: Compositional Threat Reasoning

We formalize dismantled IED assessment in X-ray imagery as a structured compositional inference problem. Unlike object-centric detection, where threat is attributed to isolated categories, we model safety as a relational property of spatially grounded components and their functional compatibility. Although Falcon-X instantiates this formulation with IED components, the same part-relation risk template can be extended to other modular threats by changing the component vocabulary and compatibility rules.

3.1 Structured Safety State

Let I denote a single-view X-ray image and $\mathcal{C} = \{\text{battery}, \text{detonator}, \text{main charge}\}$, a predefined functional component taxonomy. We define a binary presence vector $\mathbf{y} \in \{0,1\}^{|\mathcal{C}|}$, $y_c = 1$ iff at least one instance of component $c \in \mathcal{C}$ is present. To capture functional compatibility, we define a deterministic functional compatibility template $\mathbf{L} \in [0,1]^{|\mathcal{C}|\times|\mathcal{C}|}$, where L_{uv} encodes the relative functional compatibility strength between component types u and v. Higher values indicate stronger likelihood that the two components can jointly participate in a functional assembly. The matrix $\mathbf{L}$ is predefined at the type level and serves as a compatibility prior and annotation scaffold. The pair $(\mathbf{y}, \mathbf{L})$ defines a structured safety state.

3.2 Compositional Threat Formulation

Threat is treated as a relational property over present components. A minimal completeness condition is

$$\text{complete}(I) = \nVdash\{\mathbf{y} = \mathbf{1}\}, \tag{1}$$

indicating that all required component types are observed under the predefined functional taxonomy. Functional completeness captures whether the required parts are visible under the type-level template. Beyond completeness, threat depends on the compatibility among the present components. Therefore, we define a continuous scene-level risk variable $r \in [0, 1]$, representing the likelihood that grounded components admit a plausible functional assembly conditioned on $(\mathbf{y}, \mathbf{L})$.

3.3 Learning Objective

Given an image I and optional query q, the model performs grounded perception, structured state estimation, and language generation. It first predicts segmentation-aware region proposals $R = \{(b_i, m_i, s_i)\}_{i=1}^{N}$, where b_i, m_i, and s_i denote bounding box, mask, and confidence. From R, the model estimates $\hat{\mathbf{p}} \in [0, 1]^{|\mathcal{C}|}$, $\hat{\mathbf{L}} \in [0, 1]^{|\mathcal{C}|\times|\mathcal{C}|}$, $\hat{r} \in [0, 1]$, corresponding to component presence, image-conditioned functional links, and scene risk. These form a predicted safety state $(\hat{\mathbf{p}}, \hat{\mathbf{L}}, \hat{r})$. Finally, the model generates a response $Y = f(I, q)$, grounded in the inferred structured state.

4 Falcon-X Dataset

We introduce **Falcon-X**, a dual-energy X-ray benchmark for compositional threat reasoning. Unlike existing datasets focused on intact prohibited-item detection, Falcon-X models spatially dispersed functional components whose risk arises from compatibility rather than object identity (Table 1). The dataset combines dense instance grounding with structured supervision over component presence and scene-level risk, enabling evaluation of relational threat assessment.

4.1 Data Collection

Falcon-X contains 7,000 real dual-energy baggage scans with inert dismantled IED components drawn from $\mathcal{C} = \{\text{battery}, \text{detonator}, \text{main charge}\}$ [5]. Component combinations, spatial layouts, and occlusion levels are systematically varied to generate compositional configurations under realistic clutter and superposition. Additional details are provided in Appendix 9.

4.2 Dense and Structured Annotation

Each image is annotated with instance-level bounding boxes and pixel-accurate masks (Fig.3). From these annotations, we derive the structured safety state comprising: (i) component presence $\mathbf{y}$, (ii) functional completeness (Eq. 1), and (iii) Link matrix $\mathbf{L}$ defined over $\mathcal{C}$.

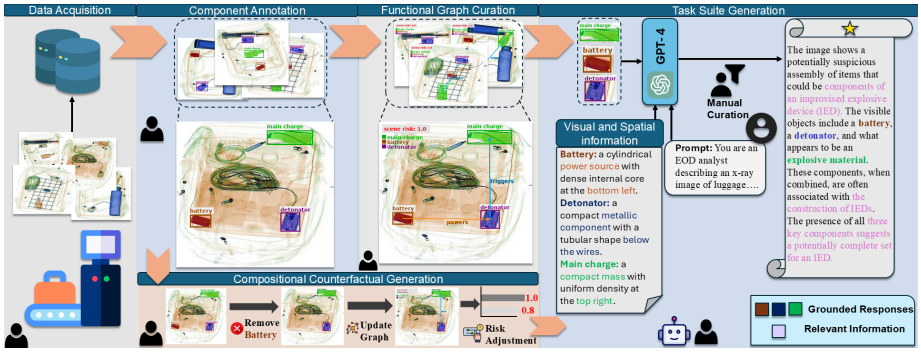

Fig. 3: Overview of the Falcon-X data collection pipeline. Collected X-ray images undergo pixel-level component-wise annotations. **Functional Graph Curation** extends those annotations into pair-wise functionality links. **Compositional Counterfactual Generation** stage applies controlled counterfactual image generation with risk supervision. **Task Suite Generation** stage automates the process of Falcon-X task suite creation followed by manual curation. 👤 indicates human involvement, while 🤖 indicates automated stages.

Each scene is additionally assigned a risk score $r \in [0, 1]$ reflecting the likelihood that the observed component configuration constitutes a functional threat. Risk labels are provided by expert annotators and incorporate perceptual uncertainty: scenes missing one component may still receive high risk due to possible occlusion or concealment. This prevents trivial derivation of r from $\mathbf{y}$ and enforces uncertainty-aware reasoning.

4.3 Counterfactual Extension

To evaluate missing-component and partial-assembly reasoning, we generate controlled counterfactual variants via mask-guided inpainting. Selected component instances are covered with a background mask obtained from the same image to preserve background clutter and semantic coherence, producing images with altered compositional states. This technique was chosen over mask zero-filling to avoid the introduction of visible artifacts. By enumerating feasible subsets of $\mathcal{C}$, we obtain a balanced corpus spanning single components, partial assemblies, and complete configurations. The final dataset comprises approximately 50,000 images (real + counterfactual). Details are provided in Appendix 9.2.

4.4 Splits and Protocol

Data are split at the base-image level (80/20 train/test), with all counterfactual variants assigned to the same partition to avoid leakage. Component distributions are preserved across splits. Annotations follow COCO format [14], augmented with structured safety labels and multimodal task definitions (Fig.3, Sec. 5).

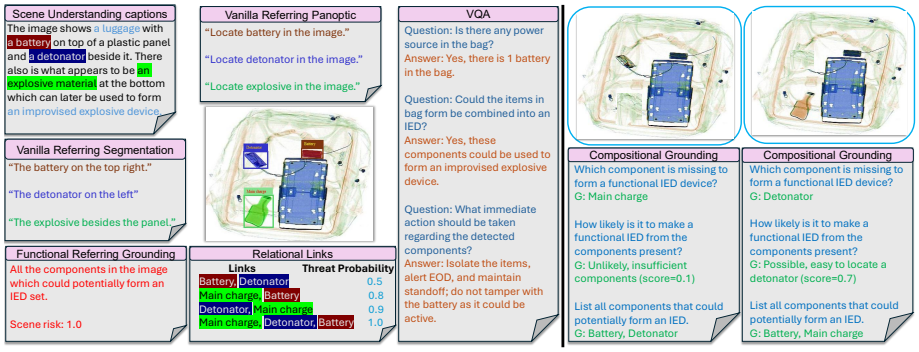

Fig. 4: Falcon-X Task Suite. Left: Example of multi-level tasks on Falcon-X base image. **Right:** Samples of compositional grounding on controlled synthetic Falcon-X images for functional reasoning.

5 Falcon-X Task Suite

Falcon-X introduces a *hierarchical evaluation framework* for compositional safety reasoning. The task suite progressively probes three levels of capability: (i) Grounded perception under X-ray superposition, (ii) compositional functional grounding, and (iii) Relationally consistent safety inference.
All tasks operate on an image I and optional query q or referring expression t. Segmentation outputs are evaluated against instance masks; language outputs are evaluated with standard captioning and QA metrics. Together, these tasks assess whether a model can construct, reason over, and coherently express a structured safety state (Fig. 4). More details on task suite generation is in Appendix Sec.10.

5.1 Layer I: Grounded perception under X-ray superposition

This layer evaluates whether a model can form an accurate perceptual state over component instances under heavy occlusion and transparency.
Scene understanding captions. The model generates a scene description of I, that identifies and enumerates visible components in $\mathcal{C}$ and describes their spatial configuration. We report BLEU, METEOR, ROUGE-L, and CIDEr.
Panoptic Segmentation. Given I, the model predicts a panoptic map assigning each pixel a semantic label (battery/detonator/charge/background) and instance IDs. We report cIoU and mIoU.
Referring Segmentation. The model predicts masks from an image I, for the referred instance. Performance is measured with cIoU and mIoU.
Referring Panoptic Segmentation. From a given t, the model predicts a panoptic map of each pixel in I belonging to the same semantic label while assigning the rest as background. Metrics of cIoU and mIoU are reported.
Visual Question Answering. We evaluate count and presence queries (e.g "How many batteries?", "Is a detonator present?"). Metrics include exact-match accuracy and MAE for counting, and binary accuracy for presence.

5.2 Layer II: Compositional Functional Reasoning

This layer evaluates whether models can reason over sets of spatially dispersed components beyond conventional object detection.
Missing Component Identification. Given a complete or partial assembly of components defined in Sec.3 and the query t "Which component is missing to form a functional IED?", the model predicts the absent class from $\mathcal{C}$. We report performance as F1 and accuracy.
Functional Completeness. The model predicts a completeness score $\hat{c} \in [0, 1]$ indicating whether the required component types are present and functionally compatible under the predefined component taxonomy. Functional completeness measures assembly sufficiency and does not directly encode expert uncertainty or contextual ambiguity. We report MAE and RMSE.
Referring functional grounding. Given a referring expression, "Ground all the components that could form a functional IED?", the model provides visually grounded regions of all possible components. Performance is measured in cIoU and mIoU.

5.3 Layer III: Relationally consistent safety inference

This layer evaluates whether models produce calibrated, relationally coherent safety assessments over the full functional state $(\mathbf{y}, \mathcal{L}, r)$.
Risk Prediction. The model predicts a continuous scene-level risk score $\hat{r} \in [0, 1]$. Risk is an expert-verified safety assessment conditioned on functional completeness, component compatibility, visual evidence, occlusion, and uncertainty. We report MAE.
Functional Link Estimation. The model predicts pairwise link probabilities $\hat{L}_{uv}$ indicating relational plausibility between component types. We evaluate using MAE for standard error assessment.
Component Set Analysis. We analyze logical consistency between predicted component presence $\hat{\mathbf{p}}$, link matrix $\mathcal{L}$, and risk $\hat{r}$ to query on grounding potential component sets. Incoherent predictions (e.g., high risk without functional completeness) are quantified to diagnose relational reasoning failures. We report set accuracy and F1 results.

6 Falcon

We propose **Falcon**, a structured segmentation-aware multimodal framework for compositional threat reasoning in X-ray imagery. Falcon introduces explicit semantic bottleneck between perception and language generation, enforcing structured safety inference prior to decoding:

$$I \rightarrow R \rightarrow \{h_c\} \rightarrow (\hat{\mathbf{p}}, \mathcal{L}, \hat{r}) \rightarrow Y,$$

where R denotes instance regions, $\{h_c\}$ component-level slot embeddings, $\hat{\mathbf{p}}$ component presence, $\mathcal{L}$ functional links, $\hat{r}$ scene-level risk, and Y the generated response. This intermediate state imposes a relational inductive bias, transforming threat assessment from implicit attention-based reasoning into structured inference.

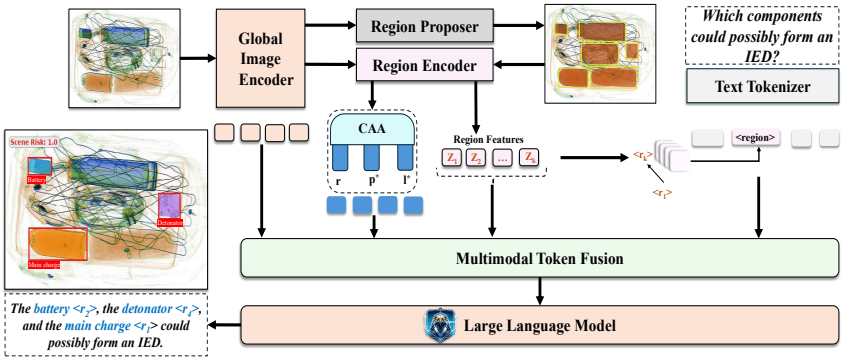

Fig. 5: Falcon architecture. Segmentation-aware perception extracts mask-aligned region embeddings, which are aggregated into structured component slots by the SSA to predict presence, functional links, and scene risk. The resulting structured tokens are fused with visual and textual tokens, introducing a relational perception before LLM decoding.

6.1 Falcon Architecture

Falcon integrates structured relational signals with region-level visual grounding to enable compositional reasoning over dismantled components. The architecture consists of four parts: (1) Segmentation-aware perception, (2) Structured functional state modeling, (3) Safety token injection, and (4) an LLM that performs multimodal reasoning and language generation. (Fig. 5)

Segmentation-aware perception. Reliable functional reasoning in X-ray imagery benefits from instance-level grounding. This is achieved by three key components:

Image encoder. We adopt DINOv2 [24] as the visual backbone and resize all X-ray inputs to 448×448. The encoder produces a dense grid of patch tokens that preserve spatial resolution and subtle contrast variations, which are critical for resolving small dismantled components under heavy clutter and translucency.

To reduce sequence length while maintaining spatial locality, non-overlapping 2×2 patches of tokens are linearly aggregated, following the token aggregation strategy adopted in [19]. This reduces the visual sequence by a factor of four while approximately preserving local structural cues. The resulting compact yet spatially structured feature map F serves as the shared visual representation for downstream instance discovery and mask-aware grounding.

Region proposer. A class-agnostic transformer detection head RF-DETR [28] extracts instance proposals $R = \{(b_i, m_i, s_i)\}_{i=1}^{N}$, including bounding boxes, segmentation masks, and objectness confidence scores. Mask prediction is critical under X-ray superposition, where bounding boxes alone capture substantial background clutter. After confidence filtering and non-maximum suppression, the top-K proposals are retained. The predicted instance masks provide precise spatial support for subsequent mask-aware region tokenization, enabling fine-grained grounding of partially occluded and fragmented components.

Mask-aware region encoder. For each proposal in R, we combine ROI-aligned features f_i^{roi}, extracted from the backbone feature map F using ROIAlign [8], with mask-pooled features obtained from the predicted segmentation mask m_i $f_i^{\text{mask}} = \frac{\sum_x m_i(x)\, F(x)}{\sum_x m_i(x)}$, where $F(x)$ denotes the backbone feature at spatial location x, through a learnable projection. The fused region embedding $z_i = W\,[f_i^{\text{roi}} \,\|\, f_i^{\text{mask}}]$ forms a language-aligned region token.

Structured functional state modeling We introduce a Structured Safety Adapter (SSA) that maps a variable-sized set of region embeddings into a fixed, interpretable functional state representation.

Query-based component slots (CAA). Let $Z = \{z_i\}_{i=1}^{N}$, where $z_i \in \mathbb{R}^d$, denote mask-aware region embeddings. SSA maintains $C = 3$ learned component queries $Q = \{q_c\}_{c=1}^{3}$, where $q_c \in \mathbb{R}^d$, each corresponding to a predefined component type in Sec.3. Each query attends to the region set via scaled dot-product attention as $A_{c,i} = \text{softmax}_i\left(\frac{q_c^\top z_i}{\sqrt{d}}\right)$. Component slot embeddings h_c are obtained from Z by weighted aggregation as $h_c = \sum_{i=1}^{N} A_{c,i} z_i$ for $c = 1, 2, 3$. This query-based aggregation produces one structured embedding per component type, allowing multiple instances to contribute proportionally. Slot specialization emerges from component-level supervision during training.

Structured Predictions. From $H = \{h_1, h_2, h_3\}$, SSA predicts three categories of structured signals:

(1) Component Presence. For each component class $c \in \mathcal{C}$, a linear projection predicts its presence probability $\hat{p}_c = \sigma(f_p(h_c))$.

(2) Pairwise Functional Links. For each predefined component pair $\mathcal{L} = \{(b, d), (b, e), (d, e)\}$, we concatenate their corresponding embeddings and a lightweight MLP head predicts their pairwise functionality link to model a graph-like association between components $\hat{\ell}_{uv} = \sigma(f_\ell([h_u; h_v])), (u, v) \in \mathcal{P}$. The link predictions model relational compatibility rather than physical connectivity.

(3) Scene Risk Level. The scene-level risk is predicted from the structured component embeddings and inferred link probabilities, rather than directly from raw visual tokens. Specifically, a lightweight MLP head estimates a scalar value $\hat{r} = \sigma\big(f_r([h_1, h_2, h_3, \hat{\ell}_{bd}, \hat{\ell}_{be}, \hat{\ell}_{de}])\big)$, $\hat{r} \in [0, 1]$.

Tokenized Safety Conditioning. Structured predictions are assembled into $v = [\hat{r}, \hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{\ell}_{bd}, \hat{\ell}_{be}, \hat{\ell}_{de}]$. Each scalar v_k is converted to a safety token $t_k = e_k^{type} + \phi(v_k)$, where e_k^{type} encodes variable identity and ϕ projects the value into the LLM embedding space. The resulting tokens are concatenated with image and region tokens prior to decoding, enabling conditioning on the inferred structured safety state.

LLM. We use Vicuna-7B [4] as the language backbone. Image, region, and safety tokens are projected into the LLM embedding space and concatenated with text tokens to form $[T_{\text{image}} \| T_{\text{region}} \| T_{\text{SSA}} \| T_{\text{text}}]$. The LLM is frozen and adapted using LoRA in attention layers for efficient domain-specific fine-tuning.

6.2 Model Training

Falcon is trained in three stages to decouple segmentation learning from structured multimodal reasoning.

Stage 1: Segmentation-Aware Proposal Learning. We train RF-DETR [28] in a class-agnostic manner to produce instance proposals and masks, treating all foreground components as a single threat class $\mathcal{L}_{S1} = \mathcal{L}_{det} + \mathcal{L}_{seg}$, where $\mathcal{L}_{det}$ denotes standard DETR set prediction losses (classification, box regression, GIoU), and $\mathcal{L}_{seg}$ combines mask BCE and Dice terms. Only RF-DETR is optimized; all multimodal modules remain frozen. The resulting proposals are fixed for subsequent stages.

Stage 2: Structured Multimodal Alignment. With RF-DETR and the LLM frozen, region proposals are converted into mask-aware embeddings Z. The Structured Safety Adapter (SSA) maps Z to component slots and predicts scene risk $\hat{r}$, presence $\hat{\mathbf{p}}$, and relational links $\hat{\mathbf{L}}$, which are injected as structured tokens into the LLM. The training objective jointly supervises language generation and structured safety prediction:

$$\mathcal{L}_{S2} = -\sum_{t} \log P_\theta(y_t \mid y_{<t}, I, Z, \text{SSA}) + \lambda_r \|\hat{r} - r\|_1 + \lambda_p \|\hat{\mathbf{p}} - \mathbf{p}\|_1 + \lambda_\ell \|\hat{\mathbf{L}} - \mathbf{L}\|_1, \quad (2)$$

Here P_θ denotes the autoregressive LLM generation process. Only the SSA and projection layers are updated.

Stage 3: Instruction Fine-Tuning. Starting from the Stage-2 checkpoint, we enable LoRA [9] adapters within the LLM while keeping RF-DETR frozen. The optimization objective remains $\mathcal{L}_{S2}$, but gradients now propagate through the LoRA parameters, improving instruction adherence and alignment between structured predictions and generated explanations.

7 Experiments

7.1 Experimental Setup

We use DINOv2-L/14 [24] as the visual backbone with 448×448 inputs. Region proposals are generated by class-agnostic RF-DETR (seg-2xlarge, 6 decoder layers) [28]. Up to 300 proposals are produced, followed by NMS (0.6) and score filtering (0.15), retaining the top 100 regions. Mask-aware region features are projected to the LLM space and aggregated by the SSA to predict risk, component presence, and pairwise links, forming seven structured safety tokens. Training follows three stages: (i) 12 epochs of detector pretraining, (ii) 1 epoch of structured multimodal alignment with frozen visual backbone and LLM, and (iii) 1 epoch of instruction tuning using LoRA [9] (rank 16). Optimization uses AdamW with mixed precision on two A100 GPUs. Additional details are provided in the Appendix Sec.13.

7.2 Results Discussion

Following Sec.5, fine-tuned models are trained and evaluated on the Falcon-X train/test splits, while zero-shot methods are evaluated on the test split only. Top three results per metric are highlighted in Red/Green/Blue, and Falcon's margin over the next best method is shown in Gray. Cross-dataset generalization results are provided in Appendix 12.

Scene Understanding and VQA. Table 2 evaluates scene understanding captioning and VQA performance of models on Falcon-X (Sec.5.1). Zero-shot general-purpose vision-language models exhibit limited transfer to X-ray imagery, reflecting the substantial distribution gap between natural images and heavily overlapped security scans. Moreover, this was also the case for domain-adapted methods like Sting-Bee [32]. After fine-tuning, however, performance saturates across all models, indicating that conventional scene understanding in this case is largely a matter of domain adaptation. Falcon performs on par with or slightly exceeding strong baselines such as Groma [15], LISA [11] and Sting-Bee [32]. Notably, the integration of structured functional state modeling preserves general multimodal capabilities while targeting higher-level compositional inference.

Table 2: Scene understanding image captioning and VQA performance on Falcon-X.

Method	Image Captioning			VQA			
	BLEU	ROUGE-L	CIDEr	BLEU	METEOR	ROUGE-L	CIDEr
Zero-shot							
Sa2Va [35]	25.06	18.11	0.012	2.88	-	3.25	-
Kosmos-2 [25]	16.26	17.56	0.002	2.83	5.47	3.79	0.003
Llava-1.5 [16]	12.67	15.67	-	2.77	7.64	3.39	-
GLaMM [27]	15.91	17.56	0.0032	1.95	4.86	3.52	0.001
LISA [11]	20.57	18.08	0.01	3.86	-	3.94	-
Groma [19]	18.8	16.48	0.004	3.32	8.90	3.44	-
Sting-Bee [32]	24.61	17.95	0.012	3.86	7.74	4.04	0.01
Fine-tuned							
LISA [11]	35.57	39.13	0.04	99.91	91.42	99.93	6.25
Sting-Bee [32]	33.38	37.48	0.024	99.12	91.27	99.92	6.24
Groma [19]	40.7	47.69	0.061	99.86	91.24	99.92	6.24
Ours	40.92 (0.22↑)	47.74 (0.05↑)	0.064 (0.003↑)	99.93 (0.02↑)	91.98 (0.56↑)	99.95 (0.02↑)	6.26 (0.01↑)

Perception and Compositional Grounding. Table 3 addresses the grounding tasks of Layers I and II evaluation in Sec.5.1 and 5.2. On conventional localization tasks (RS, PS, RPS), strong zero-shot models such as Sa2Va [35] remain competitive, and fine-tuned baselines achieve solid performance through domain adaptation. We do not consistently outperform prior methods on these appearance-driven tasks. In contrast, a distinct pattern emerges for referring functional grounding (RFG), where the target region is defined by functional composition rather than object identity. On RFG, our model outperforms the next strongest baselines by +30.38 and +40.03 points in cIoU and mIoU respectively. Notably, this relative gain is substantially larger than any difference observed on RS or RPS, indicating that the improvement does not arise from generic localization advances. The RFG improvement confirms that relational supervision enables compatibility-aware grounding. When targets are compositional rather than appearance-driven, structured modeling is essential.

Table 3: Perception and compositional grounding performance on Falcon-X. Referring segmentation, Panoptic segmentation, Referring Panoptic Segmentation, and Referring Functional Grounding are represented as RS, PS, RPS and RFG, respectively.

Method	RS		PS		RPS		RFG	
	cIoU	mIoU	cIoU	mIoU	cIoU	mIoU	cIoU	mIoU
Zero-shot								
Sa2Va [35]	14.96	32.78	35.53	54.14	17.5	37.98	14.26	20.64
Kosmos-2 [25]	2.95	3.2	2.78	3.08	2.92	3.32	3.61	3.62
GLaMM [27]	12.37	19.84	6.41	13.72	11.28	18.93	5.73	17.02
LISA [11]	3.50	6.73	9.13	14.21	11.50	16.95	11.86	13.24
Groma [19]	11.61	20.49	7.88	16.04	14.06	21.2	8.16	26.31
Sting-Bee [32]	15.23	22.37	12.45	15.69	13.21	14.95	10.45	11.89
Fine-tuned								
LISA [11]	11.82	14.75	35.18	37.12	30.12	33.34	9.45	12.98
Sting-Bee [32]	18.72	26.81	31.05	33.49	15.40	19.59	12.72	14.96
Groma [19]	14.25	21.99	36.73	13.78	17.94	24.1	20.07	29.55
Ours	25.86 (7.14↑)	35.39 (2.61↑)	14.37 (21.16↓)	38.02 (16.12↓)	21.74 (8.38↓)	35.30(2.68↓)	50.45 (30.38↑)	69.58 (40.03↑)

Compositional Threat Reasoning. Table 4 reports semantic grounding (Layer II) and relational safety metrics (Layer III). Zero-shot MLLMs perform poorly on compositional tasks, particularly MCI and relational safety measures, indicating limited transfer from object-level semantics to functional reasoning. After fine-tuning, component presence (CPC) saturates across models (≈98%), suggesting that independent component recognition is not the main challenge. In contrast, relational metrics remain substantially harder. Accurate estimation of functional completeness (FC), scene risk (SRL), and link risk (CLR) requires compatibility-aware reasoning rather than isolated detection. Falcon achieves competitive CPC while consistently reducing MAE on FC, SRL, and especially CLR (0.005), indicating improved relational calibration. The gains concentrate on compatibility-dependent metrics despite near-saturated presence scores, demonstrating that improvements arise from structured modeling of component interactions rather than improved object classification.

Table 4: Compositional Semantic Grounding and Safety Relation analysis performance on Falcon-X. Component Presence Check, Missing Component Identification, Functional Completeness, Scene Risk Level, Potential Component Sets, and Component Link Risk are represented as CPC, MCI, FC, SRL, PCS and CLR, respectively.

Method	Semantic Grounding						Safety Relationship			
	CPC		MCI		FC		SRL	PCS		CLR
	Acc↑	F1↑	Acc↑	F1↑	MAE↓	RMSE↓	MAE↓	Acc↑	F1↑	MAE↓
Zero-shot										
Sa2Va [35]	58.4	46.2	12.6	15.8	0.60	–	0.40	13.56	54.9	–
Kosmos-2 [25]	57.09	36.36	14.16	15.26	0.57	0.67	0.43	14.3	57.1	–
Llava-1.5 [16]	57.4	47.12	14.32	14.93	0.7	–	–	14.1	56.47	–
GLaMM [27]	55.8	35.9	11.6	14.2	0.52	–	0.31	12.6	56.2	–
LISA [11]	–	–	14.29	14.29	0.9	0.9	–	–	–	–
Groma [19]	57.1	36.4	13.3	19.9	0.59	0.69	0.36	14.63	57.05	–
Sting-Bee [32]	57.14	36.36	14.29	14.29	0.49	0.48	0.42	0.01	42.86	–
Fine-tuned										
LISA [11]	-	-	17.63	19.25	0.28	0.35	0.20	0.41	43.06	0.20
Sting-Bee [32]	98	98	93.45	96.31	0.031	0.36	0.23	14.89	56.34	0.24
Groma [19]	98	98	97.1	98.6	0.019	0.13	-	14.9	57.06	-
Ours	98.1 (0.1↑)	97.97 (0.03↓)	94.75 (2.35↓)	97.33 (1.27↓)	0.017 (0.002↓)	0.09 (0.04↓)	0.02 (0.18↓)	15.1 (0.2↑)	57.69 (0.59↑)	0.005 (0.195↓)

7.3 Qualitative Results

Fig. 6 shows qualitative examples of referring functional grounding and scene-level reasoning. Falcon localizes components that jointly satisfy functional assembly criteria, including spatially separated and partially occluded instances. For incomplete assemblies, it identifies missing components and adjusts risk pre-

dictions consistently with the inferred structured state. Additional examples are provided in Appendix 11.

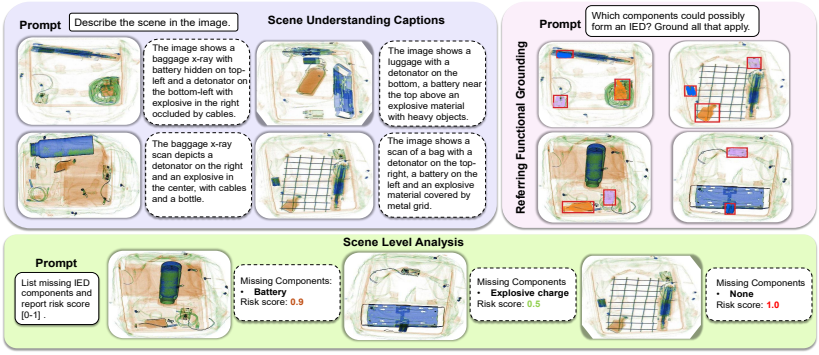

Fig. 6: Falcon qualitative results on compositional grounding and scene-level reasoning.

7.4 Ablations

We conduct controlled ablations to quantify the contribution of (i) structured prediction heads and (ii) perception versus reasoning bottlenecks. All variants are evaluated on the primary task of *Referring Functional Grounding* (RFG). Additional ablations are reported in Appendix 14.

Prediction head ablations. We ablate the Structured Safety Adapter (SSA) by selectively disabling the risk, presence, and link heads prior to token injection. As shown in Table 5, removing all heads reduces Falcon to a region-grounded VLM without structured signals. Enabling the risk head alone yields limited improvement, while adding presence prediction provides a larger gain. The most significant increase occurs when the link head is activated, indicating that relational compatibility modeling is the primary driver of functional grounding performance.

Oracle ablations. Introduced to decouple the relative contributions of perception and structured reasoning modules (Table 6). *Oracle-Reasoning* replaces the predicted functional selection with ground-truth relational assignments over the same predicted regions, isolating the decision layer under fixed perception. *Oracle-Perception* replaces predicted regions with ground-truth masks while retaining learned reasoning, providing a perception upper bound. The gap Falcon→Oracle-reasoning measures reasoning headroom, while Oracle-reasoning→Oracle-perception quantifies the remaining perception bottleneck.

Table 5: SSA Prediction heads ablations on referring functional grounding (RFG). R, P, and L denote the risk, component-presence, and link prediction heads. Performance is reported in cIoU and mIoU.

Variant			RFG	
R	P	L	cIoU	mIoU
✗	✗	✗	35.82	43.43
✓	✗	✗	39.33 (3.51↑)	47.98 (4.55↑)
✓	✓	✗	45.52 (6.19↑)	58.60 (10.62↑)
✓	✓	✓	**50.45** (4.93↑)	**69.58** (10.98↑)

Table 6: Oracle ablation performance of Falcon perception and reasoning stages on referring functional grounding (RFG). Results are reported in cIoU and mIoU and Δvs. Falcon

Mode	Perception	Reasoning	RFG	
			cIoU	mIoU
Falcon	Pred	Pred	50.45	69.58
Oracle-Reasoning	Pred	GT	59.61 Δ(9.16↑)	79.53 Δ(9.95↑)
Oracle-Perception	GT	Pred	79.49 Δ(29.04↑)	83.42 Δ(14.84↑)

8 Conclusion

We formulate functional threat reasoning in X-ray imagery as a compositional inference problem, where safety emerges from the structured interaction of spatially dispersed components rather than isolated object detection. To address this setting, we introduce Falcon, a segmentation-aware multimodal framework that injects explicit structured relational state between perception and language generation, and Falcon-X, a benchmark that jointly evaluates grounding, compositional reasoning, and calibrated risk assessment. Through controlled experiments and diagnostic ablations, we show that explicit relational modeling substantially improves functional grounding and safety-consistent prediction under heavy superposition. By coupling structured supervision with multimodal reasoning, Falcon establishes a principled approach to compositional safety analysis, while Falcon-X provides a unified evaluation protocol for this emerging problem. We hope this work motivates further research on structured, safety-critical reasoning in complex visual environments.

Acknowledgements

This research was supported by Khalifa University Digital Future Institute (KU-DF), and the Khalifa University Center for Autonomous Robotic Systems (KU-CARS).

References

1. Cai, Z., Ke, F., Jahangard, S., Garcia de la Banda, M., Haffari, R., Stuckey, P.J., Rezatofighi, H.: Naver: A neuro-symbolic compositional automaton for visual grounding with explicit logic reasoning. In: ICCV (2025)
2. Chang, A., Zhang, Y., Zhang, S., Zhong, L., Zhang, L.: Detecting prohibited objects with physical size constraint from cluttered x-ray baggage images. Knowledge-Based Systems (2022)
3. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023)
4. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), `https://lmsys.org/blog/2023-03-30-vicuna/`
5. Eshetu, A., Burton, T.B., Howell, J.D., Rutter, M.F., Winnett, T.J.: Inert ied training kits. United States Patent Application US20160161228A1 (June 2016)
6. Garcia-Fernandez, P., Vaquero, L., Liu, M., Xue, F., Cores, D., Sebe, N., Mucientes, M., Ricci, E.: Superpowering open-vocabulary object detectors for x-ray vision. In: ICCV (2025)
7. He, C., Mu, T., Ren, W., Zhao, B.: Lpixray: A large-scale logistics prohibited item x-ray dataset for the application of deep learning in security inspection. CIPAE (2023)
8. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017)

9. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)
10. Lafon, M., Karmim, Y., Silva-Rodríguez, J., Couairon, P., Rambour, C., Fournier-Sniehotta, R., Ayed, I.B., Dolz, J., Thome, N.: Vilu: Learning vision-language uncertainties for failure prediction. In: ICCV (2025)
11. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: CVPR (2024)
12. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
13. Li, M., Jia, T., Wang, H., Ma, B., Lu, H., Lin, S., Cai, D., Chen, D.: Ao-detr: Anti-overlapping detr for x-ray prohibited items detection. TNNLS **36** (2024)
14. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
15. Liu, C., Ding, H., Jiang, X.: GRES: Generalized referring expression segmentation. In: CVPR (2023)
16. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
17. Ma, B., Jia, T., Li, M., Wu, S., Wang, H., Chen, D.: Toward dual-view x-ray baggage inspection: A large-scale benchmark and adaptive hierarchical cross refinement for prohibited item discovery. ITIFS **19**, 3866–3878 (2024)
18. Ma, B., Jia, T., Su, M., Jia, X., Chen, D., Zhang, Y.: Automated segmentation of prohibited items in x-ray baggage images using dense de-overlap attention snake. ITMM **25**, 4374–4386 (2022)
19. Ma, C., Jiang, Y., Wu, J., Yuan, Z., Qi, X.: Groma: Localized visual tokenization for grounding multimodal large language models. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 417–435. Springer Nature Switzerland, Cham (2025)
20. Mao, J., Gan, C., Kohli, P., Tenenbaum, J.B., Wu, J.: The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In: ICLR (2019)
21. Mery, D., Riffo, V., Zscherpel, U., Mondragón, G., Lillo, I., Zuccar, I., Lobel, H., Carrasco, M.: Gdxray: The database of x-ray images for nondestructive testing. Journal of Nondestructive Evaluation (2015)
22. Miao, C., Xie, L., Wan, F., Su, c., Liu, H., Jiao, j., Ye, Q.: Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images. In: CVPR (2019)
23. Michael, Y., Alansari, M., Ahmed, A., Werghi, N., Henschel, A.: X-ssl: Self-supervised x-ray threat detection with zero-shot and multi-modal learning. Information Processing & Management **63**(1) (2026)
24. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
25. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. ArXiv **abs/2306.14824** (2023)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)

27. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. CVPR (2024)
28. Robinson, I., Robicheaux, P., Popov, M., Ramanan, D., Peri, N.: RF-DETR: Neural architecture search for real-time detection transformers. In: ICLR (2026)
29. Tao, R., Li, H., Wang, T., Wei, Y., Ding, Y., Bowei Jin and, H.Z., Liu, X., Liu, A.: Exploring endogenous shift for cross-domain detection: A large-scale benchmark and perturbation suppression network. In: CVPR (2022)
30. Tao, R., Wei, Y., Jiang, X., Li, H., Qin, H., Wang, J., Ma, Y., Zhang, L., Liu, X.: Towards real-world x-ray security inspection: A high-quality benchmark and lateral inhibition module for prohibited items detection. In: ICCV (2021)
31. Tao, R., Wei, Y., Jiang, X., Li, H., Qin, H., Wang, J., Ma, Y., Zhang, L., Liu*, X.: Towards real-world x-ray security inspection: A high-quality benchmark and lateral inhibition module for prohibited items detection. In: ICCV (2021)
32. Velayudhan, D., Ahmed, A., Alansari, M., Gour, N., Behouch, A., Hassan, T., Wasim, S.T., Maalej, N., Naseer, M., Gall, J., et al.: Sting-bee: Towards vision-language model for real-world x-ray baggage security inspection. In: CVPR (2025)
33. Velayudhan, D., Hassan, T., Damiani, E., Werghi, N.: Recent advances in baggage threat detection: A comprehensive and systematic survey. ACM Computing Surveys **55**(8) (2022)
34. Wei, Y., Tao, R., Wu, Z., Ma, Y., Zhang, L., Liu, X.: Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module. In: ACMMM (2020)
35. Yuan, H., Li, X., Zhang, T., Huang, Z., Xu, S., Ji, S., Tong, Y., Qi, L., Feng, J., Yang, M.H.: Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. arXiv (2025)
36. Zhang, L., Jiang, L., Ji, R., Fan, H.: Pidray: A large-scale x-ray benchmark for real-world prohibited item detection. IJCV (2023)
37. Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Liu, Y., Chen, K., Luo, P.: Gpt4roi: Instruction tuning large language model on region-of-interest. In: Del Bue, A., Canton, C., Pont-Tuset, J., Tommasi, T. (eds.) Computer Vision – ECCV 2024 Workshops. pp. 52–70. Springer Nature Switzerland, Cham (2025)
38. Zhao, C., Zhu, L., Dou, S., Deng, W., Wang, L.: Detecting overlapped objects in x-ray security imagery by a label-aware mechanism. ITIFS **17**, 998–1009 (2022)
39. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. In: ICLR (2024)