Can a Second-View Image Be a Language? Geometric and Semantic Cross-Modal Reasoning for X-ray Prohibited Item Detection

Chuang Peng[1], Renshuai Tao[1]*, Zhongwei Ren[1], Xianglong Liu[2], Yunchao Wei[1]
[1]Institute of Information Science, Beijing Jiaotong University
[2]State Key Laboratory of Complex & Critical Software Environment, Beihang University
mr.pengc@foxmail.com, rstao@bjtu.edu.cn

Abstract

Automatic X-ray prohibited items detection is vital for security inspection and has been widely studied. Traditional methods rely on visual modality, often struggling with complex threats. While recent studies incorporate language to guide single-view images, human inspectors typically use dual-view images in practice. This raises the question: ***can the second view provide constraints similar to a language modality?*** *In this work, we introduce DualXrayBench, the first comprehensive benchmark for X-ray inspection that includes multiple views and modalities. It supports eight tasks designed to test cross-view reasoning. In DualXrayBench, we introduce a caption corpus consisting of 45,613 dual-view image pairs across 12 categories with corresponding captions. Building upon these data, we propose the* ***G****eometric (cross-view)-****S****emantic (cross-modality)* ***R****easoner (GSR), a multimodal model that jointly learns correspondences between cross-view geometry and cross-modal semantics, treating the second-view images as a "language-like modality". To enable this, we construct the GSXray dataset, with structured Chain-of-Thought sequences: <top>, <side>, <conclusion>. Comprehensive evaluations on DualXrayBench demonstrate that GSR achieves significant improvements across all X-ray tasks, offering a new perspective for real-world X-ray inspection.*

1. Introduction

Automatic X-ray prohibited items detection [25, 26, 32, 33, 41] is crucial for security inspection and has been extensively studied. Traditional methods [1, 23, 38] focus on extracting low-level visual features, such as edges and contours, to detect prohibited items. However, these systems struggle with real-world scenarios due to limited generalization and lack of semantic reasoning, making them unable to handle emerging threats or structural variations, such as distinguishing tablets from laptops [26]. As a result, purely visual-modal-based models fail to provide robust detection.

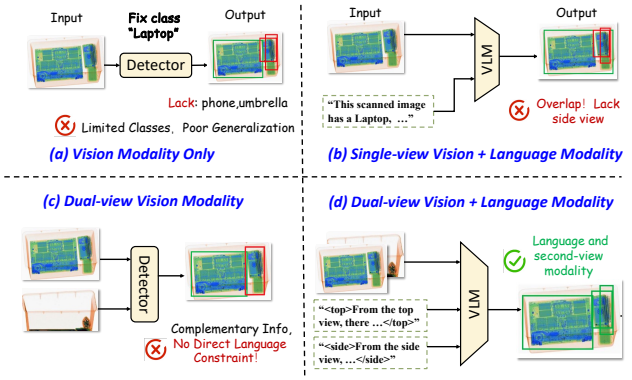

Figure 1. Comparison of existing X-ray prohibited-item detection strategies (a–c) with our DualXrayBench (d), which treats the second-view image as a language-like modality that provides additional constraints to enhance detection.

The limitations of vision-only systems have sparked interest in integrating semantic reasoning through language modalities, driven by advancements in Vision-Language Models (VLMs) [20, 29, 40]. This multi-modal approach enhances reasoning and generalization by grounding visual data in semantic space, allowing models to infer beyond pixel-level patterns. Recent studies have leveraged language to overcome the closed-set limitations of vision-only X-ray systems, using image-caption pairs and textual descriptors to regularize visual learning [17, 37]. While language-grounded models effectively address intra-class variance and improve open-vocabulary detection, dual-view X-ray imagery, a crucial modality used by human inspectors, has been overlooked. Human inspectors rely on dual views to recognize prohibited items despite occlusions and structural ambiguities. Inspired by the idea that "images are also language", we had a sudden insight: **Can a second-view image be treated as a language, with its geometric correspondences jointly learned with linguistic semantics to enhance X-ray prohibited items detection?**

*Corresponding author.

Realizing this joint reasoning paradigm requires datasets that simultaneously capture semantic, geometric, and visual correspondences, but resources are largely absent in the X-ray domain. Existing single-view X-ray datasets enriched with image-caption pairs [17, 37] provide valuable semantic grounding but lack the geometric complementarity needed for cross-view reasoning. In contrast, emerging X-ray dual-view datasets [35, 42] capture structural consistency between perspectives but remain vision-only, lacking linguistic alignment for semantic learning. As a result, no unified resource bridges visual, geometric, and linguistic modalities, limiting Vision-Language Models' ability to leverage the inherent dual-view consistency of X-ray imagery and generalizing in real-world security applications.

To address these challenges, in this work, we introduce **DualXrayBench**, the first comprehensive benchmark for X-ray prohibited items detection that integrates multiple views and modalities. DualXrayBench includes a dual-view caption corpus with 45,613 pairs of images across 12 categories, each accompanied by captions. Each sample provides: (1) image-level scene analyses for individual top and side views, (2) a holistic dual-view description combining complementary information, and (3) fine-grained object-level annotations with cross-view associations, including category, spatial relations, and occlusion relations. Moreover, DualXrayBench supports eight tasks designed to test cross-view reasoning, comprising 1,594 expert-verified visual question–answer pairs. Unlike conventional single-view evaluation settings, DualXrayBench employs a dual-view evaluation paradigm, assessing models' reasoning ability, cross-view consistency, and spatial understanding across eight diagnostic sub-tasks.

Building upon these data, we propose the **G**eometric-**S**emantic **R**easoner (**GSR**), a multimodal model that jointly learns textual semantics and dual-view geometric correspondences, treating the second-view image as a structured "language-like modality". To enable this, we specifically construct the **GSXray dataset** with structured Chain-of-Thought (CoT) sequences: <top>, <side>, <conclusion>, explicitly modeling the reasoning flow from view-specific observations to unified semantic conclusions. Evaluations on DualXrayBench demonstrate that GSR achieves significant and consistent improvements across all tasks, showcasing superior cross-view reasoning, compositional understanding, and robustness to unseen prohibited items categories. The contributions are summarized as follows:

- We are the first to propose that a second-view image can be treated as a language-like modality in X-ray detection task, providing extra constraints to enhance performance.
- We introduce DualXrayBench, the first benchmark for X-ray detection with cross-view and cross-modal data, supporting eight tasks to assess cross-view reasoning, interpretability, consistency, and spatial understanding.
- We propose the GSR method, a multimodal model that jointly learns correspondences between cross-view geometry and cross-modal semantics, achieving significant improvements across the eight X-ray detection tasks.

2. Related Work

X-ray Baggage Security Datasets. Recent advances in computer-aided X-ray screening have been largely driven by deep learning and the release of public datasets [25, 26, 32, 38, 41] such as SIXray [26], PIDray [38], and HiXray [32], which have propelled research on object detection and segmentation of prohibited items, as summarized in Table 1. Although these datasets comprise hundreds of thousands of scans and address challenges like class imbalance and occlusion, most adhere to a single-view, closed-set paradigm. Crucially, in real-world security inspection, human operators routinely examine both top and side view scans to mentally reconstruct 3D object structures and resolve occlusion-induced ambiguities. This motivates dual-view research [3, 24, 31, 35, 42], exemplified by datasets such as DvXray [42], which provide paired scans but remain constrained by

Dataset	2-view	Caption	Open	N_c	N_p	Task	Year
GDXray [25]	✗	✗	✓	3	8,150	D	2015
Compass XP [4]	✗	✗	✓	366	1928	C	2019
SIXray [26]	✗	✗	✓	6	8,929	C, D	2019
OPIXray [41]	✗	✗	✓	5	8,885	D	2020
PIDray [38]	✗	✗	✓	12	47,677	D, S	2021
HiXray [32]	✗	✗	✓	8	45,364	D	2021
PIXray [23]	✗	✗	✓	15	5,046	D, S	2022
CLCXray [46]	✗	✗	✓	12	9,565	D	2022
EDS [33]	✗	✗	✓	10	14,219	D	2022
FSOD [34]	✗	✗	✓	20	12,333	D	2022
LPIXray [9]	✗	✗	✗	18	60,950	D	2023
MV-Xray [31]	✓	✗	✗	2	3,231	D	2019
deei6 [3]	✓	✗	✗	6	7,022	D/S	2021
DBF [11]	✓	✗	✗	4	2,528	D	2021
Dualray [42]	✓	✗	✗	6	4,371	D	2022
DvXray [24]	✓	✗	✓	15	5,000	C	2024
LDXray [35]	✓	✗	✓	12	146,997	D	2024
STCray [37]	✗	✓	✓	21	46,642	D/S/ZS	2024
PIXray Caption [17]	✗	✓	✓	15	5,046	D/ZS	2025
DualXrayBench	✓	✓	✓	12	45,613	D/ZS	2025

Table 1. A comprehensive overview of X-ray datasets used in security inspection, detailing the number of categories (N_c) and images containing prohibited items (N_p). The tasks include Classification (C), Detection (D), Segmentation (S), and Zero-Shot (ZS).

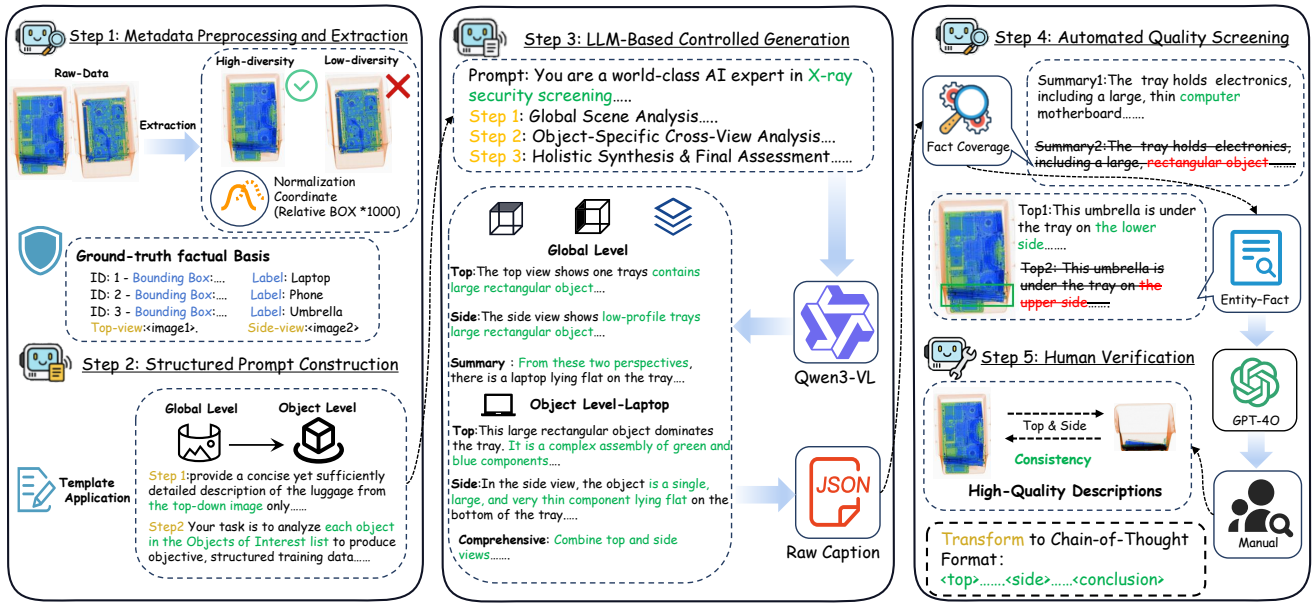

Figure 2. Construction pipeline of the DualXrayBench caption corpus, illustrating metadata preprocessing and extraction, structured prompt design, LLM-based controlled caption generation, automated quality filtering, and expert verification to ensure high-quality captions for X-ray prohibited-item detection and reasoning.

scale and annotation granularity. Existing dual-view models primarily rely on feature fusion [10, 35], treating the auxiliary view as additional imagery without explicitly modeling cross-perspective visibility or geometric consistency key factors for human-level spatial reasoning in dual-view X-ray analysis

Vision-Language Models in X-ray Security. The success of large Vision-Language Models (VLMs) [16, 20, 40], exemplified by foundational models like GLIP [16], has opened new avenues for open-vocabulary detection and semantic understanding. Recent research has adapted VLMs to specialized domains [15], including medical imaging and X-ray security. Efforts like OVXD [18] adapts CLIP for open-vocabulary X-ray detection but relies solely on word-level textual supervision, limiting its semantic understanding. OVPID introduces the PIXray-Caption dataset [17] with 5,046 image–caption pairs to alleviate the closed-set limitation, yet its limited caption quantity and coarse image-level descriptions constrain fine-grained visual–language alignment. STCray [37] brings instruction-following to X-ray analysis via text–image pairs but remains restricted to single-view alignment without modeling cross-view geometry. Overall, existing work advances X-ray vision–language modeling yet still lacks fine-grained data and explicit cross-view reasoning.

3. DualXrayBench

To enable structured cross-perspective visual–language reasoning, we introduce **DualXrayBench**, including a large-scale hierarchical caption corpus built upon the dual-view LDXray dataset [35]. LDXray contains 146,997 paired top and side view X-ray images with annotated bounding boxes and categories. Based on these, we generate 45,613 verified dual-view caption pairs capturing both scene-level and fine-grained cross-view correspondences across 12 classes: Mobile Phone (MP), Orange Liquid (OL), Portable Charger 1, lithium-ion prismatic cell (PC1), Portable Charger 2, lithium-ion cylindrical cell (PC2), Laptop (LA), Green Liquid (GL), Tablet (TA), Blue Liquid (BL), Columnar Orange Liquid (CO), Nonmetallic Lighter (NL), Umbrella (UM) and Columnar Green Liquid (CG).

3.1. Corpus Construction Pipeline

Pipeline Overview. As illustrated in Fig. 2, the caption corpus is built through a semi-automated, five-stage pipeline ensuring factual correctness, linguistic coherence, and cross-view consistency.

(1) Preprocessing. We filter low-diversity samples and normalize object bounding boxes to a $[0, 1000]$ scale, producing resolution-independent structured metadata comprising category and spatial information.

(2) Structured Prompting. Each image pair and its metadata are formatted into a hierarchical prompt instructing the model to describe the top and side view scenes, provide object-level analyses, and summarize global semantics under strict JSON constraints to avoid hallucination.

(3) LLM-Based Generation. Using Qwen3-VL-235B-A22B-Thinking [36], we generate hierarchical captions including global layout, per-object reasoning, and holistic summaries. Descriptions explicitly encode complementary cross-view cues (e.g., "flat in top-view but vertically ex-

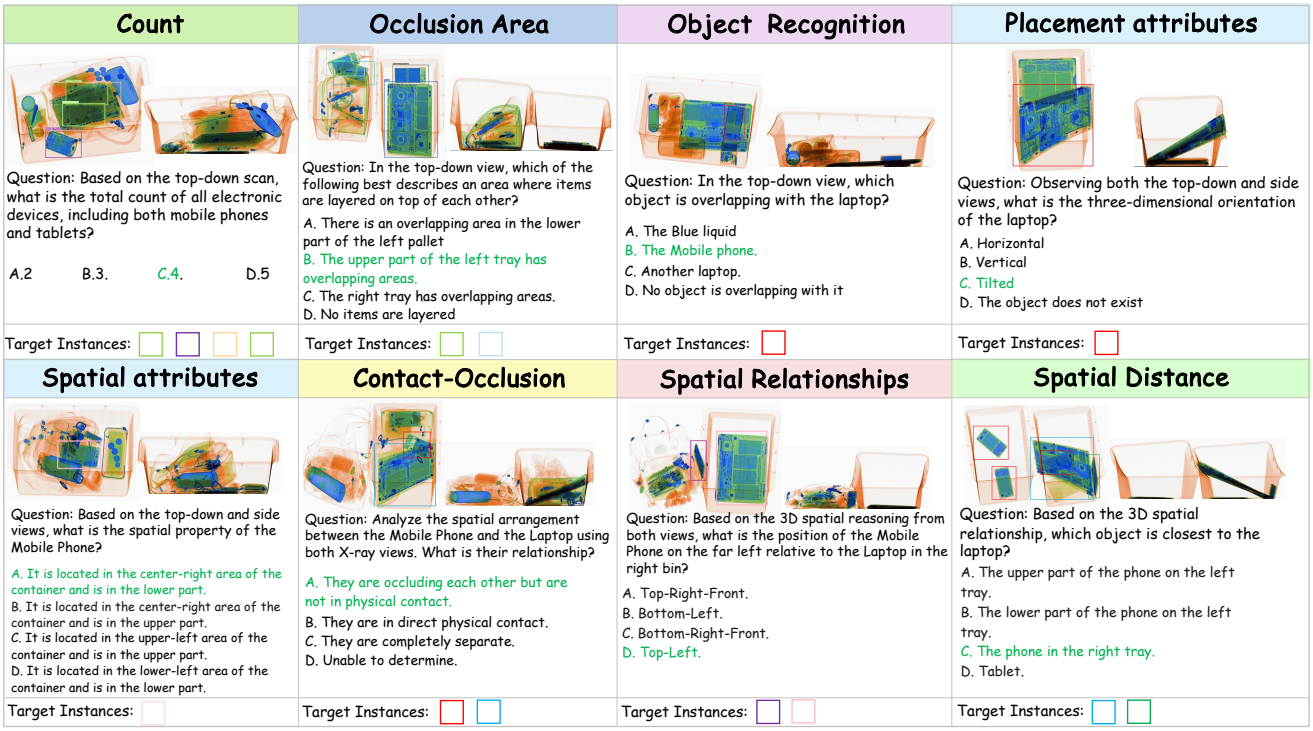

Figure 3. Representative examples from DualXrayBench illustrating eight diagnostic tasks. Each example pairs top- and side-view images with annotated evidence and corresponding questions. Tasks encompass perception (counting, recognition), relational reasoning (spatial alignment), occlusion inference (visibility and containment), and attribute understanding (posture, geometry), highlighting the benchmark's coverage of cross-perspective reasoning challenges.

tended in side-view").

(4) Automated Screening. A three-stage validation ensures quality through: (i) fact coverage, (ii) entity–fact alignment with bounding boxes, and (iii) cross-view consistency checking via GPT-4o [27]. Samples failing coverage ($< 80\%$), alignment ($< 50\%$), and consistency ($> 50\%$ contradiction) are discarded.

(5) Human Verification. Domain experts conduct multi-round review to ensure spatial and semantic accuracy, revising inconsistent or ambiguous cases. The verified corpus exhibits high factual precision and naturalness.

Resulting Corpus. The final Corpus provides rich, hierarchical dual-view captions suitable for visual–language reasoning and instruction tuning. Furthermore, its structured form supports the construction of the **GSXray dataset** by converting captions into a Chain-of-Thought format <top>,<side>,<conclusion> with the assistance of GPT-4o [27]. This design explicitly supervises cross-perspective reasoning alignment in multimodal models. **Details are provided in the Supplementary Materials.**

3.1.1. Data Statistics

Category and Instance Distributions. DualXrayCap contains 45,613 paired images with a total of 139,956 instances spanning 12 common categories, as summarized in Table 2. The dataset exhibits diverse annotation patterns, with 5,211 images containing a single instance and 38,967 images having multiple instances, resulting in an average of 3.07 annotations per image. Moreover, 22,772 images contain overlapping objects with an Intersection over Minimum (IoM) greater than 0.3, indicating frequent occlusions and complex spatial arrangements.

Category	MP	OL	PC1	PC2	LA	GL	TA	BL	CO	NL	UM	CG	Total
Number	59,003	18,684	9,998	17,274	13,453	11,482	6,796	1,677	513	487	298	296	139,956

Table 2. Category-wise statistics of DualXray corpus.

3.2. DualXrayBench Construction Pipeline

We introduce **DualXrayBench**, a multi-task benchmark comprising 1,594 expert-verified visual question–answer pairs. It is constructed through a semi-automated pipeline that converts structured image–text data from the DualXray corpus into a comprehensive suite of measurable reasoning tasks, with strict checks to prevent overlap with the GSXray dataset. The benchmark is produced via a three-stage process that integrates automated generation with expert validation to ensure reliability in cross-perspective visual–language reasoning.

Stage I Source corpus preparation We randomly sample 5,000 instances from the DualXray corpus, which provides top and side view scene analyses, fine-grained object-level descriptions, and holistic summaries. From these, we

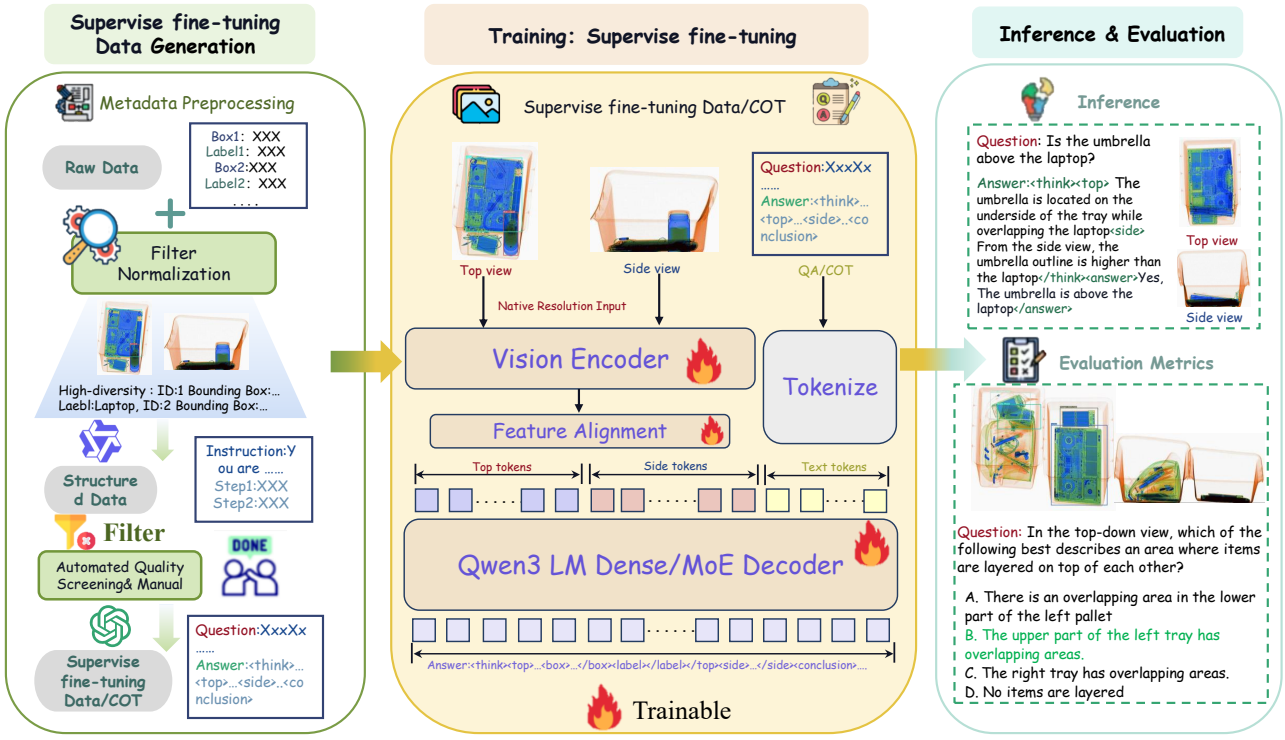

Figure 4. Overall architecture of the DualXrayBench–GSR framework: (1) A structured data generation pipeline for creating DualXrayBench and GSXray CoT supervision from raw multi-view X-ray metadata; (2) A supervised fine-tuning framework that learns dual-view geometric correspondences and cross-modal semantics, treating side-view imagery as a language-like modality; (3) A unified inference and evaluation protocol for cross-view reasoning, consistency assessment, and spatial understanding across eight diagnostic tasks.

extract structured scene facts including object categories, spatial relationships, and occlusion and containment relations to serve as verified inputs for question generation.

Stage II Multi-task Q&A generation Each sample is used to generate four distinct question–answer pairs via controlled LLM prompting, corresponding to: (1) *Perceptual* (counting/identification), (2) *Relational* (spatial reasoning), (3) *Occlusion* (visibility inference) and (4) *Attribute* (pose/property recognition). Each includes a reference answer and traceable evidence (bounding boxes and labels).

Stage III Human review & difficulty calibration All samples undergo two-stage review: expert quality control for clarity and answerability, followed by model-consensus filtering to remove trivial or overly hard cases. The final benchmark contains 1,594 high-quality, human-verified samples, ensuring both difficulty balance and traceable evaluation integrity.

3.2.1. Task Definition

To comprehensively evaluate cross-perspective reasoning, **DualXrayBench** defines four task families *Perceptual*, *Relational*, *Occlusion*, and *Attribute* each comprising representative sub-tasks that emphasize dual-view understanding (Fig. 3).

Perceptual Tasks. This category includes *Counting* (CT) and *Object Recognition* (OR). Unlike conventional single-view setups, both tasks are constructed under occlusion conditions to highlight the complementary role of side-view information. CT evaluates a model's ability to count partially hidden instances, while OR examines cross-view category identification consistency.

Relational Tasks. We design *Spatial Relation* (SR) and *Spatial Distance* (SD) tasks to assess geometric reasoning across perspectives. SR requires inferring the relative position of object pairs (e.g., above/below, front/behind), while SD focuses on estimating inter-object distances, particularly along the depth (z)-axis challenging for single view perception.

Occlusion Tasks. This family comprises *Occluded Area Recognition* (OA) and *Contact–Occlusion Judgment* (CO). OA targets identifying occluded regions by integrating complementary visibility cues across views, while CO requires reasoning about whether two objects are physically contacting or mutually occluding demanding explicit cross-view inference.

Attribute Tasks. Finally, *Placement Attribute* (PA) and *Spatial Attribute* (SA) tasks evaluate the model's understanding of object posture and spatial configuration. The former focuses on pose level properties (e.g., standing, flat, tilted), and the latter examines localization in 3D space, especially along the vertical dimension.

Together, these eight sub-tasks form a structured, multi-faceted evaluation protocol that probes not only recognition accuracy but also the depth of geometric and causal reasoning enabled by dual-view perception, with the number of instances per sub-task summarized in the Table 3.

Task	CT	OR	SR	SD	OA	CO	PA	SA	Total
Number	200	200	200	200	196	199	199	200	1594

Table 3. Sample distribution of the eight diagnostic sub-tasks.

4. Geometric-Semantic Reasoner (GSR)

Building upon these data, we propose the GSR, a multimodal model that jointly learns textual semantics and dual-view geometric correspondences, treating the second-view image as a structured "language-like modality".

4.1. GSXray dataset

To enable supervision for cross-view geometric semantic reasoning, we construct the **G**eometric–**S**emantic dual-view X-ray dataset (**GSXray**), which comprises 44,019 dual-view QA samples organized into structured Chain-of-Thought sequences `<top>`, `<side>`, and `<conclusion>`. Unlike conventional SFT CoT datasets, the **GSXray** designed for dual view X-ray inspection decomposes reasoning into view specific observations that are fused into a unified semantic conclusion. This structure offers fine-grained supervision for modeling the shift from geometric perception to high-level semantics, akin to how human inspectors integrate top side views to resolve occlusions. These CoT sequences are automatically transformed from DualXray Corpus with lightweight GPT assistance and minimal manual oversight. Representative examples are illustrated in the figure 4. **More details are provided in the Supplementary Materials.**

4.2. GSR Implementation

To enable fine-grained cross-perspective understanding in X-ray imagery, we propose the **Geometric–Semantic Reasoner (GSR)** a multimodal reasoning architecture built upon the Qwen3-VL-MoE-8B [36] foundation model. GSR extends the visual–language alignment capability of large multimodal models (LMMs) by explicitly encoding geometric correspondence and semantic hierarchy across dual views. As shown in figure 4, the architecture comprises three key components: a vision encoder $\boldsymbol{E}$, a Feature Alignment $\boldsymbol{A}$, and a language reasoning model $\boldsymbol{L}$.

The vision encoder $\boldsymbol{E}$ follows the ViT-L/14 design within the Qwen3-VL-MoE framework. It independently processes the paired dual-view X-ray inputs top view $\boldsymbol{x}_{\text{top}}$ and side-view $\boldsymbol{x}_{\text{side}}$ through a shared-weight transformer backbone equipped with 3D convolutional patch embedding and rotary positional encoding. The encoder extracts dense feature representations that preserve both geometric structures and spatial context, producing:

$$\boldsymbol{f}_i = \boldsymbol{E}(\boldsymbol{x}_i) \in \mathbb{R}^{m \times n}, \tag{1}$$

where $\boldsymbol{f}_i$ denotes a sequence of vision tokens encoding object-level and contextual cues from the top and side view image. The multi-level attention mechanism within the MoE transformer allows for efficient feature aggregation across scales, capturing view-dependent relationships essential for X-ray scene interpretation.

A lightweight multimodal alignment module $\boldsymbol{A}$ implemented as a two-layer MLP with GeLU activation maps visual embeddings into the LLM's semantic space:

$$\boldsymbol{f}'_i = \boldsymbol{A}(\boldsymbol{f}_i). \tag{2}$$

This alignment bridges the geometric representation space of the vision encoder and the linguistic embedding space of the LLM, producing the projected feature tokens $\boldsymbol{f}'_i \in \mathbb{R}^k$. These tokens are then fed into the LLM as structured multimodal inputs.

Hierarchical Reasoning Tokens. Unlike previous architectures [2, 14, 48] that treat visual embeddings as untyped tokens, GSR introduces **hierarchical reasoning tokens** `<top>`, `<side>` and `<conclusion>` to explicitly encode **cross-view semantics** [5, 13, 15, 37]. The `<top>` and `<side>` tokens distinguish visual evidence originating from orthogonal viewpoints, guiding the model to maintain spatial awareness and disambiguate occluded structures. The `<conclusion>` token serves as an aggregation signal, prompting the LLM to integrate geometric and semantic cues into a unified reasoning output. These tokens enable the model to dynamically adjust its reasoning granularity—from fine-grained threat localization to scene-level compositional inference—depending on the user query and the structural hierarchy implied by the token context.

Language Reasoning. The **language reasoning module** $\boldsymbol{L}$, instantiated by the Qwen3-VL-MoE text decoder, jointly processes the projected visual tokens $\boldsymbol{f}'_i$ and the textual query sequence $\boldsymbol{t}$. Through multi-head attention and mixture-of-experts routing, the model generates responses conditioned on both the dual-view geometric cues and task-specific semantics. This process can be summarized as:

$$\boldsymbol{y} = \boldsymbol{L}\big([\texttt{<top>}\ \boldsymbol{f}'_{\text{top}},\ \texttt{<side>}\ \boldsymbol{f}'_{\text{side}},\ \texttt{<conclusion>}\ \boldsymbol{t}]\big), \tag{3}$$

where $\boldsymbol{y}$ denotes the generated reasoning output.

By explicitly incorporating geometric semantic reasoning tags, GSR overcomes a key limitation of prior multimodal models—the inability to differentiate tasks requiring distinct spatial granularity. The structured dual-view representation allows GSR to perform coherent cross-view reasoning, supporting diverse tasks such as cross-view correspondence analysis, spatial relationship inference, and multi-view consistency verification.

Method	Overall			Perception		Relational		Occlusion		Attribute	
	Acc	F1	mIOU	CT	OR	SR	SD	OA	CO	PA	SA
GPT-4o [27]	47.0	49.2	16.5	46.5	45.5	43.0	38.5	43.4	57.8	47.5	59.8
GPT-o3 [28]	49.3	52.6	19.7	47.1	48.9	50.5	36.4	43.5	59.2	52.5	56.3
Gemini-2.5-Flash [7]	45.1	48.9	14.7	43.0	42.9	41.2	34.5	42.8	56.0	51.0	49.4
Gemini-2.5-Pro [8]	58.6	60.5	28.7	56.3	43.0	44.3	61.5	52.9	74.6	64.3	71.9
Qwen2.5-VL-7B [2]	50.3	53.0	20.3	35.5	43.5	47.5	48.5	43.4	52.8	65.3	65.8
Qwen2.5-VL-32B [2]	53.0	55.4	24.5	45.0	42.5	47.0	54.0	50.0	53.8	58.1	74.0
Qwen2.5-VL-72B [2]	57.1	59.5	26.3	54.0	43.7	39.5	59.0	52.6	72.4	65.6	70.5
Qwen3-VL-4B [36]	47.2	50.2	20.2	18.2	41.5	45.0	46.5	46.9	53.3	52.9	59.5
Qwen3-VL-8B [36]	53.5	56.6	25.4	45.5	42.9	38.0	47.0	54.6	71.9	62.6	66.5
Qwen3-VL-30B [36]	51.8	53.3	25.4	49.2	47.0	46.0	54.5	50.5	51.3	53.3	62.5
Qwen3-VL-235B [36]	58.8	65.5	26.0	54.0	55.8	40.5	**61.0**	56.6	69.3	54.0	**79.0**
InternVL2-8B [6]	33.8	37.4	8.4	29.9	35.5	29.9	31.6	30.8	39.2	36.7	36.4
InternVL2.5-8B [6]	36.3	40.6	10.2	27.3	36.9	30.8	32.5	38.8	43.9	43.3	37.2
InternVL3-8B [48]	43.6	45.4	15.3	34.3	32.1	28.6	34.8	44.0	57.2	56.8	61.0
InternVL3-30B [48]	45.4	47.3	15.7	32.5	40.1	29.7	36.5	46.0	58.4	60.3	59.7
InternVL3-78B [48]	48.9	51.6	16.4	43.5	39.1	28.5	41.2	48.0	65.8	63.9	61.2
InternVL3.5-8B [39]	45.4	51.5	18.1	43.7	29.5	28.0	36.5	49.0	61.8	57.6	58.0
InternVL3.5-30B [39]	46.1	49.3	18.5	37.5	41.5	30.5	37.5	46.4	60.9	61.8	54.0
InternVL3.5-241B [39]	51.3	53.8	22.0	46.8	43.0	30.2	42.5	49.6	73.8	**66.2**	58.3
LLaVA1.5-7B [20]	15.4	20.5	6.2	19.5	11.0	16.0	9.5	15.3	21.6	12.6	18.0
LLaVA-Next-7B [22]	24.5	28.3	13.1	20.6	27.1	24.0	23.1	20.2	27.3	27.8	25.4
LLaVA-Next-34B [22]	28.3	36.5	14.5	24.0	31.7	26.0	26.3	24.7	38.2	28.1	27.4
LLaVA-Next-110B [22]	33.6	38.8	16.8	27.0	25.1	24.5	31.5	31.8	53.0	40.0	36.0
LLaVA-OneVision-7B [19]	32.3	36.8	15.6	34.5	33.0	28.5	33.7	29.1	30.8	31.5	37.0
LLaVA-OneVision-72B [19]	42.2	45.1	19.2	34.5	40.0	40.0	41.5	44.9	45.2	47.7	43.7
STING-BEE [19]	23.8	29.8	13.2	22.6	26.4	23.6	24.1	19.4	23.4	24.5	26.8
Qwen2.5-VL-7B(Ours)	55.1 ↑4.8%	62.2 ↑9.2%	28.7 ↑8.4%	46.2 ↑10.7%	45.3 ↑1.8%	47.9 ↑0.4%	55.7↑7.2%	51.6 ↑8.2%	64.5 ↑11.7%	66.1↑0.8%	63.5
InternVL3.5-8B(Ours)	50.3 ↑4.9%	54.5 ↑3.0%	23.2 ↑5.1%	52.0 ↑8.3%	44.0 ↑14.5%	24.5	44.0 ↑7.5%	52.0 ↑3%	73.9 ↑12.1%	52.3	60.0 ↑2.0%
LLaVA-Next-7B(Ours)	29.5 ↑5.0%	34.0 ↑5.7%	23.1 ↑10.0%	19.0	37.6 ↑10.5%	**54.1**↑1.1%	33.2 ↑10.1%	26.2 ↑6.0%	28.9 ↑1.6%	26.6	38.8 ↑13.4%
LLaVA-OneVision-7B(Ours)	39.3 ↑7.0%	47.2 ↑10.4%	28.3 ↑12.7%	38.2 ↑3.7%	37.6 ↑4.6%	30.1↑1.6%	42.6 ↑8.9%	37.3 ↑8.2%	45.7 ↑14.9%	43.1↑11.6%	39.8 ↑2.8%
GSR-8B (Ours)	**65.4**↑11.9%	**70.6**↑14%	**52.3** ↑26.9%	**63.5**↑18%	**65.8** ↑22.9%	44.0 ↑6%	**61.0**↑14%	**67.9** ↑13.3%	**80.4** ↑8.5%	64.4↑1.8%	76.5 ↑10%

Table 4. Comparison with state-of-the-art alternatives on DualXrayBench. All results are self-collected and each row's **Overall** equals the arithmetic mean of the eight task columns (CT, OR, SR, SD, OA, CO, PA, SA). The best performance is highlighted in **bold**.

5. Experiments

Implementation. We initialize our Geometric–Semantic Reasoner (GSR) from the Qwen3-VL-8B[36] foundation checkpoint. Both the visual encoder and the multimodal alignment module are fully unfrozen during training to enable end-to-end optimization. Fine-tuning is performed using LLaMA-Factory with mixed-precision (bfloat16) on eight NVIDIA H200 GPUs. We employ the AdamW optimizer with a base learning rate of 1e-6, cosine decay scheduling and a warmup ratio of 0.1. The global batch size is set to 256, and the model is trained for two epochs.

5.1. DualXrayBench Evaluation

Evaluation Metrics. We evaluate model performance across the eight dual-view reasoning tasks using three metrics: mean Intersection-over-Union (mIoU), Accuracy (Acc), and F1-score. mIoU assesses spatial correspondence, Accuracy measures prediction correctness and F1-score captures the balance between precision and recall, jointly reflecting the model's geometric and semantic reasoning ability.

Main Results. We comprehensively evaluate state-of-the-art VLMs on **DualXrayBench**, covering both proprietary and open-source systems. As shown in Table 4, most off-the-shelf models struggle with dual-view X-ray understanding, with overall accuracy typically below 50. For example, GPT-4o [27] achieves only 47.0 accuracy, and InternVL3.5-8B [39] reaches 45.4, reflecting their limited generalization in this domain. Notably, STING-BEE, a single-view X-ray VLM, performs even worse at 23.8, underscoring its inability to handle dual-view reasoning. In contrast, our **GSR-8B** establishes a new state-of-the-art with 65.4 accuracy and 52.3 mIoU, outperforming all baselines by a large margin. It consistently leads across perception and reasoning subtasks, such as 63.5 in CT and 61.0 in SD.

Model	Top view	<Top>	Side view	<Side>	Acc	F1	mIoU
Qwen3vl-8B [36]	✗	✗	✗	✗	53.5	56.6	25.4
+ <answer>	✓	✗	✗	✗	56.8↑ 3.3	58.3↑ 3.3	25.2
	✓	✓	✗	✗	59.2↑ 5.7	63.7↑ 7.1	41.1↑ 15.7
	✓	✗	✓	✗	57.3↑ 3.8	60.5↑ 3.9	24.6
	✓	✓	✓	✗	62.1↑ 8.6	65.8↑ 9.2	45.2↑ 19.8
	✓	✗	✓	✓	58.6↑ 5.1	61.3↑ 4.7	25.7↑ 0.5
	✓	✓	✓	✓	**65.4**↑ 11.9	**70.6**↑ 14	**52.3**↑ 26.9

Table 5. Ablation of dual-view and reasoning components.

Impact of GSXray Fine-tuning. To further validate the effectiveness of the GSXray dataset, we fine-tuned representative open-source VLMs, including Qwen2.5-VL-7B [2], InternVL3.5-8B [39], and LLaVA-Next-7B [22]. All models exhibit substantial and consistent improvements across perception, relational, and occlusion tasks. Notably, LLaVA-OneVision-7B improves from 32.2 to 39.3 accu-

Model	GSXray	Dual view	Acc	F1	mIOU
Qwen2.5-VL-7B [2]	✗	✗	49.4	52.6	22.6
		✓	50.3 ↑0.9%	53.0 ↑0.4%	20.3 ↓2.3%
InternVL3.5-8B [39]	✗	✗	43.2	50.3	18.6
		✓	45.4 ↑2.2%	51.5 ↑1.0%	18.1 ↓0.5%
LLaVA-OneVision-7B [19]	✗	✗	33.1	38.5	18.2
		✓	32.3 ↓0.8%	36.8 ↓1.7%	15.6 ↓2.6%
InternVL3.5-8B (Ours)	✓	✗	45.2	49.6	20.5
		✓	50.3 ↑5.1%	54.5 ↑4.9%	23.2 ↑2.7%
LLaVA-OneVision-7B (Ours)	✓	✗	37.8	45.7	27.3
		✓	39.3 ↑1.5%	47.2 ↑1.5%	28.3 ↑1.0%
GSR-8B (Ours)	✓	✗	62.1	67.4	49.6
		✓	65.4 ↑3.3%	70.6 ↑3.2%	52.3 ↑2.7%

Table 6. Ablation studies on the effect of the second view.

racy, and InternVL3.5-8B rises from 45.4 to 50.3, revealing remarkable performance gains. These results indicate that GSXray transfers domain knowledge effectively, enabling coherent cross-view geometric–semantic reasoning and substantially reducing the domain gap.

5.2. Ablation Study

Second View and Structured Reasoning. To evaluate the contributions of the second view and structured reasoning, we perform controlled ablations on DualXrayBench using Qwen3-VL-8B (Table 5). Using only the primary (top) view yields limited gains, highlighting that single-view perception is insufficient for spatial reasoning. Adding structured top-view descriptions (`<top>` with `<box>` grounding) improves performance, especially mIoU. Including the side view further enhances perceptual consistency and relational understanding, while dual views without textual guidance offer only marginal benefits and may reduce mIoU due to missing grounding. Full structured reasoning (`<top>`, `<side>`, `<conclusion>`, `<answer>`) with dual views achieves the strongest gains, forming GSR-8B. These results indicate that visual complementarity and structured reasoning are both essential, jointly enabling precise geometric alignment and robust cross-view inference.

Role of the second view on DualXrayBench. Table 6 analyzes the influence of incorporating the second view on DualXrayBench before and after GSXray fine-tuning. Before training, the additional view yields inconsistent gains sometimes even degrading performance suggesting that vanilla VLMs fail to exploit cross-view cues, as unaligned side-view features act more as noise than complementary evidence. After fine-tuning, this behavior flips: removing the second view consistently harms Acc, F1, and mIoU. This shift verifies that GSXray effectively learns to extract and align semantic signals from the auxiliary view, converting an underutilized modality into a structured constraint that reinforces geometric grounding and cross-view reasoning.

5.3. Generalization

STING-BEE-VQA Evaluation. We further benchmark GSR-8B on the STING-BEE [37] VQA benchmark, which evaluates multimodal reasoning across seven categories: *Instance Location* (IL), *Complex Reasoning* (CR), *Instance Identity* (ID), *Instance Counting* (IC), *Misleading* (ML), *Instance Attribute* (IA), and *Instance Interaction* (IT). We fine-tune GSR-8B on 10k SFT data of STING-BEE and evaluate under the same multiple-choice format. As reported in Table 7, GSR-8B consistently surpasses all vision-language baselines, showing the largest improvements in IL (+3.3), IC (+1.8), and CR (+3.8). These gains highlight stronger geometric–semantic reasoning and robustness to occlusion, showing that explicit geometric–semantic alignment enhances generalization across X-ray domains.

Model	IL	CR	ID	IC	ML	IA	IT	Overall
Florence-2 [44]	30.1	37.5	39.8	30	21.1	35.8	29.1	32.3
MiniGPT [5]	43.2	60.2	54.8	28.2	19.6	33.0	29.7	36.6
LLaVA 1.5 [21]	29.7	56.7	74.0	34.2	13.8	41.0	24.0	42.0
STING-BEE [37]	49.2	79.2	80.0	45.2	27.8	53.0	35.0	52.8
GSR-8B (Ours)	**52.5**	**83.0**	**83.5**	**47.0**	**29.5**	**57.0**	**41.1**	**55.3**

Table 7. VQA performance comparison on **STING-BEE**.

Model	Scissors	Wrench	Pliers	Battery	**AP50**
OV-DETR [45]	0.2	0.1	0.1	0.0	0.1
RegionCLIP [47]	0.4	2.6	2.4	0.0	1.3
BARON [43]	1.5	18.1	3.1	0.6	37.8
Faster RCNN [30]	0.0	0.2	0.1	0.0	0.1
YOLO11 [12]	1.2	0.8	0.2	0.1	2.3
OVXD [18]	16.9	46.2	6.4	14.6	21.0
GSR-8B (Ours)	24.2	50.5	10.1	20.4	26.3

Table 8. Cross-category and cross-domain generalization on **PID** following the OVXD protocol.

Cross-Category and Domain Generalization. Following OVXD [18], we evaluate GSR-8B on PID under cross-category and cross-domain splits. *Battery* (aligned with *Power bank* in DualXrayCap) measures domain robustness, while the remaining unseen classes test category-level generalization. Under the standardized OVXD protocol for both VLMs and vision-only baselines, GSR-8B achieves the highest accuracy across all settings, confirming that structured cross-view reasoning and geometric–semantic alignment markedly improve transferability to novel categories and scanner shifts.

6. Conclusion

In this work, we introduced DualXrayBench, the first comprehensive benchmark for X-ray inspection incorporating dual-view images and multimodal data, designed to evaluate cross-view reasoning. We presented a dual-view caption corpus and the GSXray dataset, which introduces structured Chain-of-Thought sequences to enable joint learning of cross-view geometry and cross-modal semantics. By leveraging these datasets, we proposed the GSR, which treats the second-view images as a "language-like modality". Extensive evaluations on DualXrayBench show that GSR significantly improves performance across X-ray tasks, offering a new sight for real-world X-ray inspection applications.

References

[1] Abdelfatah Ahmed, Divya Velayudhan, Taimur Hassan, Mohammed Bennamoun, Ernesto Damiani, and Naoufel Werghi. Enhancing security in x-ray baggage scans: A contour-driven learning approach for abnormality classification and instance segmentation. *Engineering Applications of Artificial Intelligence*, 130:107639, 2024. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6, 7, 8

[3] Neelanjan Bhowmik, Yona Falinie A Gaus, and Toby P Breckon. On the impact of using x-ray energy response imagery for object detection via convolutional neural networks. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1224–1228. IEEE, 2021. 2

[4] Matthew Caldwell and Lewis D Griffin. Limits on transfer learning from photographic image data to x-ray threat detection. *Journal of X-ray Science and Technology*, 27(6): 1007–1020, 2019. 2

[5] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 6, 8

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 7

[7] Google DeepMind. Gemini-2.5-flash. https://deepmind.google/models/gemini/flash/, 2025. 7

[8] Google DeepMind. Gemini-2.5-pro. https://deepmind.google/models/gemini/pro/, 2025. 7

[9] Chengquan He, Tong Mu, Weiping Ren, and Bohua Zhao. Lpixray: A large-scale logistics prohibited item x-ray dataset for the application of deep learning in security inspection. In *2023 International Conference on Computers, Information Processing and Advanced Education (CIPAE)*, pages 481–485, 2023. 2

[10] Shilong Hong, Yanzhou Zhou, and Weichao Xu. Dagnet: A dual-view attention-guided network for efficient x-ray security inspection. *arXiv preprint arXiv:2502.01710*, 2025. 3

[11] Brian K. S. Isaac-Medina, Chris G. Willcocks, and Toby P. Breckon. Multi-view object detection using epipolar constraints within cluttered x-ray security imagery. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9889–9896, 2021. 2

[12] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. 8

[13] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27831–27840, 2024. 6

[14] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6

[15] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 6

[16] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 3

[17] Shuyang Lin, Tong Jia, Hao Wang, Bowen Ma, and Mingyuan Li. Open-vocabulary prohibited item detection for real-world x-ray security inspection. *IEEE Transactions on Information Forensics and Security*, 20:7469–7481, 2025. 1, 2, 3

[18] Shuyang Lin, Tong Jia, Hao Wang, Bowen Ma, Mingyuan Li, and Dongyue Chen. Detection of novel prohibited item categories for real-world security inspection. *Engineering Applications of Artificial Intelligence*, 144:110110, 2025. 3, 8

[19] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 7, 8

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1, 3, 7

[21] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 8

[22] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 7

[23] 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. *IEEE Transactions on Multimedia*, 2022. 1, 2

[24] 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 Transactions on Information Forensics and Security*, 2024. 2

[25] Domingo Mery, Vladimir Riffo, Uwe Zscherpel, German Mondragón, Iván Lillo, Irene Zuccar, Hans Lobel, and Miguel Carrasco. Gdxray: The database of x-ray images for nondestructive testing. *Journal of Nondestructive Evaluation*, 34(4):42, 2015. 1, 2

[26] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2119–2128, 2019. 1, 2

[27] OpenAI. Openai-gpt-4o. https://openai.com/index/gpt-4o-system-card/, 2024. 4, 7

[28] OpenAI. Openai-o3. https://openai.com/index/introducing-o3-and-o4-mini/, 2025. 7

[29] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[30] Shaoqing Ren, Kaiming He, Ross B Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2017. 8

[31] Jan-Martin O Steitz, Faraz Saeedan, and Stefan Roth. Multi-view x-ray r-cnn. In *German Conference on Pattern Recognition*, pages 153–168. Springer, 2018. 2

[32] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10923–10932, 2021. 1, 2

[33] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21157–21167. IEEE, 2022. 1, 2

[34] 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 *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2012–2020, 2022. 2

[35] Renshuai Tao, Haoyu Wang, Yuzhe Guo, Hairong Chen, Li Zhang, Xianglong Liu, Yunchao Wei, and Yao Zhao. Dual-view x-ray detection: Can ai detect prohibited items from dual-view x-ray images like humans? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10338–10347, 2025. 2, 3

[36] Qwen Team. Qwen3 technical report, 2025. 3, 6, 7

[37] Divya Velayudhan, Abdelfatah Ahmed, Mohamad Alansari, Neha Gour, Abderaouf Behouch, Taimur Hassan, Syed Talal Wasim, Nabil Maalej, Muzammal Naseer, Juergen Gall, et al. Sting-bee: Towards vision-language model for real-world x-ray baggage security inspection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20767–20777, 2025. 1, 2, 3, 6, 8

[38] Boying Wang, Libo Zhang, Longyin Wen, Xianglong Liu, and Yanjun Wu. Towards real-world prohibited item detection: A large-scale x-ray benchmark. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5412–5421, 2021. 1, 2

[39] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 7, 8

[40] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Videogrounding-dino: Towards open-vocabulary spatio-temporal video grounding. In *CVPR*, 2024. 1, 3

[41] Yanlu Wei, Renshuai Tao, Zhangjie Wu, Yuqing Ma, Libo Zhang, and Xianglong Liu. Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module. In *Proceedings of the 28th ACM international conference on multimedia*, pages 138–146, 2020. 1, 2

[42] Modi Wu, Feifan Yi, Haigang Zhang, Xinyu Ouyang, and Jinfeng Yang. Dualray: Dual-view x-ray security inspection benchmark and fusion detection framework. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 721–734. Springer, 2022. 2

[43] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15254–15264, 2023. 8

[44] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024. 8

[45] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European conference on computer vision*, pages 106–122. Springer, 2022. 8

[46] 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 transactions on information forensics and security*, 17:998–1009, 2022. 2

[47] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16793–16803, 2022. 8

[48] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 6, 7