STING-BEE : Towards Vision-Language Model for Real-World X-ray Baggage Security Inspection

Supplementary Material

A. Additional details on STING Protocol

The STING protocol underpins the STCray dataset, categorizing clutter into four levels—Limited, Medium, Heavy, and Extreme— progressively increasing occlusions and distractions. Concealment sublevels further diversify scenarios, ranging from low-density (e.g., organic items like books) to extreme configurations such as metallic grids and multi-layered superimposed materials. These concealments are further diversified by systematically varying the position and orientation of the threat, distorting threat appearances, and challenging model detection capabilities. Fig. 1 illustrates this interplay, showcasing the *Scissors* across clutter, concealment, and positional variations.

Caption Generation Using the STING Protocol The caption generation process, outlined in algorithm 1, leverages prior knowledge of threat-specific metadata and synonym sets to dynamically construct captions for each X-ray image $I \in D_x$ collected using the STING Protocol. Synonym pools $S_s = \{S_x, S_p, S_t, S_c\}$ provide linguistic variations for X-ray descriptors (S_x), positioning phrases (S_p), threat descriptions (S_t), and concealment details (S_c). For each image I, attributes such as threat type $T(I)$, concealment level $L_c(I)$, location $L_l(I)$, and orientation $\Phi(I)$ are extracted to generate diverse and descriptive captions C_g, enriching the STCray dataset for vision-language tasks.

Please note that our STING protocol utilizes predetermined metadata (e.g., threat type, placement, orientation) based on input from airport security personnel. Then it generates detailed captions directly from the metadata recorded during baggage preparation and scanning, eliminating human error. To validate the annotations, two independent annotators review a subset using the metadata and X-ray scans. We obtained a ROUGE-L value of 0.7 for captions, validating the annotations.

Algorithm 1: Caption Generation using STING.

Input: Threat items T, Concealment levels L_c, Locations L_l, Synonym sets S_s, X-ray scans D_x
Output: Generated captions C_g

Initialize synonym pools S_s for X-ray descriptors, verbs, positions, and concealment phrases;
Parse threat-specific metadata and mapping rules from D_x;

foreach *image* $I \in D_x$ **do**

Extract:
- Threat item $T(I)$;
- Concealment level $L_c(I)$;
- Location $L_l(I)$ (center or corner);
- Orientation $\Phi(I)$ (e.g., horizontal, tilted);
- Concealment phrase based on $L_c(I)$.

Select a random synonym for:
- X-ray descriptor from S_s;
- Positioning verb from S_s;
- Location phrase from S_s.

Construct Caption:

Combine elements into the caption:
`"{X-ray descriptor} {positioning phrase} a {threat}, {concealment details}, {position}.";`

Append constructed caption to C_g;

end

return C_g;

B. Additional details on Instruction Tuning dataset

The Visual Question-Answering (VQA) instructions were generated using GPT-4 [4] for training STING BEE. These conversations, derived from STCray training set captions, focused on threat identification, complex reasoning, and positional queries. A carefully designed GPT prompt (Fig. 3) guided the generation process, ensuring task-specific queries and precise responses. Fig. 5 illustrates an example of a multi-round VQA conversation paired with its associated baggage scan and caption.

C. Additional details on VQA Evaluation

To evaluate STING-BEE's visual reasoning capabilities, we developed a comprehensive VQA dataset, drawing inspira-

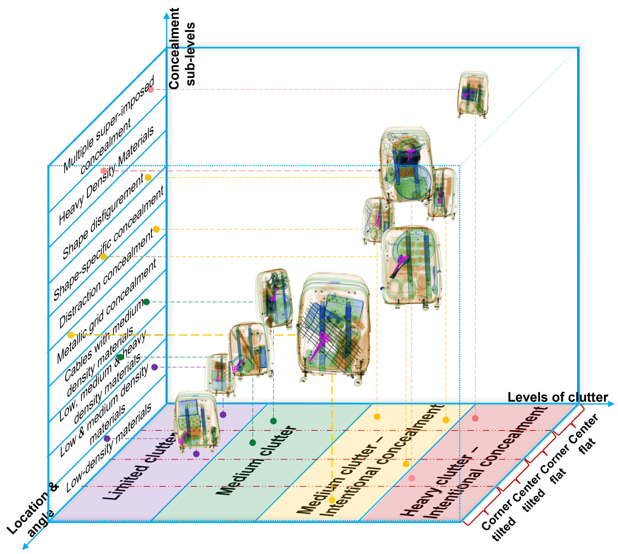

Figure 1. 3D representation of STING protocol displaying the interplay between clutter levels, concealment sublevels, and location (from central to corner and flat to inclined) using *Scissors* (with one sample for each sublevel for clarity).

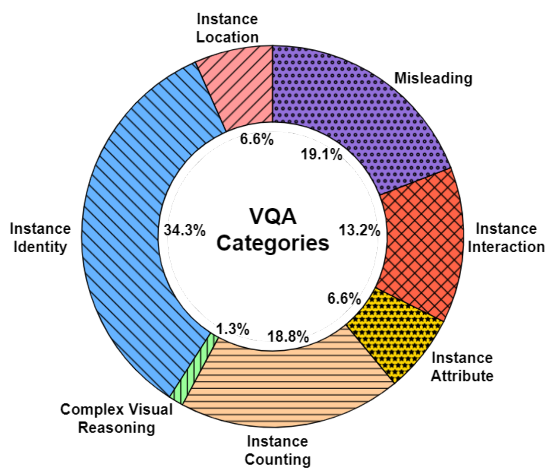

Figure 2. **VQA Evaluation Benchmark:** Distribution across seven categories to assess model robustness across diverse visual reasoning dimensions.

tion from the SEED Bench [11] methodology. It comprises 39,194 questions derived from STCray, SIXray, and PIDray. These questions span seven distinct categories, as shown in Fig. 2, targeting critical dimensions of visual reasoning.

- **Instance Identity**: Tests the model's ability to recognize and classify specific threat types (e.g., gun vs. knife), requiring image-wide context.
- **Instance Counting**: Evaluates the model's capacity to count potential threats in a scene, crucial for understanding scenario complexity.
- **Instance Location**: Assesses spatial understanding by determining object locations (e.g., center vs. corner).
- **Instance Attribute**: Focuses on identifying object-specific features like orientation or occlusion.
- **Instance Interaction**: Requires object-level reasoning and examines how objects interact, such as occlusion.
- **Complex Visual Reasoning**: Requires the model to infer threats from contextual cues, emphasizing domain-specific higher-order reasoning.
- **Misleading Questions**: Includes deliberately misleading queries to evaluate model precision.

Sample questions from each category are presented in Table 1.

Table 1. **Question categories and sample questions:** Overview of the seven question categories used in the Visual Question Answering (VQA) dataset, comprising 39,194 questions across STCray, SIXray, and PIDray.

Evaluation Dimensions	Sample Question-Answer Pairs
1. Instance Identity	What type of threat item is present in this X-ray image? A. Pliers B. Injection C. Lighter D. Battery
2. Instance Location	Where is the Battery located in this baggage X-ray scan? A. Corner B. Middle C. Not Present
3. Instance Interaction	How is the Battery concealed in this baggage X-ray scan? A. Fully Covered B. Half Covered C. Uncovered
4. Instance Attribute	In what orientation is the Battery positioned within the baggage X-ray scan? A.Vertically B. Horizontally C. At an inclined angle.
5. Instance Counting	How many potential threats are present in this X-ray image? A. One B. Two C. Three D. None
6. Complex Visual Reasoning	How can you infer that there is a metal gun in the baggage scan? A. Since the gun is bluish or greenish, it can be easily inferred as metal gun. B. Since the gun is orange-coloured, it can be easily inferred as metal gun. C. None of the above.
7. Misleading	Where is the Cutter located in the baggage scan? A. There is no cutter in the image. B. Toward the corner of the image. C. In the middle of the image.

D. Additional details on Referring threat localization and Visual Grounding

STING-BEE's performance in visual grounding and referring threat localization was evaluated using SIXray and PIDray datasets. In SIXray, which includes five threat categories, we grouped threats into two material groups: *Metallic* (*Gun*, *pliers*, *wrench*) and *Sharp* (*Knife*, *Scissors*) to enable a more granular evaluation of model performance across different material types. While PIDray's categories were organized into five groups: *Metallic*, *Sharp*, *Flammable*, *Corrosive*, and *Explosive*. Evaluation questions followed

Prompt for VQA Instruction Generation:

You are an AI assistant analyzing X-ray baggage scans to detect prohibited items and security threats. Based on a description of the scan, answer questions as if you are visually analyzing the image. The description includes objects present in the scan, potential threat items, and objects placed to conceal them. Metallic items, such as guns, knives, and pliers, appear blue; organic items, such as 3D-printed guns and improvised explosives, appear orange; and inorganic items, such as circuits, powerbank, and battery, appear green. Using the description of the scan, design a conversation between you and a person asking about this scan, focusing on identifying threat items concealed within normal items.

The following are the threat categories likely to be present in the image alongside normal items: explosive, gun, 3D-printed gun, knife, bullet, syringe, battery, wrench, other sharp items, powerbank, scissors, hammer, pliers, and screwdriver. If none of the threat items are present, and only normal items are detected, the image is classified as "Nonthreat." Note that explosives can be intact or dispersed (dismantled). If dispersed, the description will mention the positions or concealment of the three main parts of the explosive: the container with explosive material, the circuit, and the battery. Sometimes the circuit, container, or battery may be expertly concealed within normal items.

Additionally, note that 3D-printed guns are difficult to detect because of their faint outlines, polymer-based structure, and orange appearance in the scan. You can include misleading questions about threat items that are not present and answer confidently that they are not present. Furthermore, tangled wires, cables, chains, stacked metallic items, circuits, and laptops may appear suspicious in the description. You can incorporate questions to clarify if there are any suspicious items in the image. Provide confident and definite answers, avoiding any uncertain or speculative responses.

Figure 3. Prompt used for generating VQA instructions in STING-BEE. The prompt guides GPT-4 to generate conversations focusing on identifying concealed threat items in X-ray baggage scans, as if you are visually analyzing the scan.

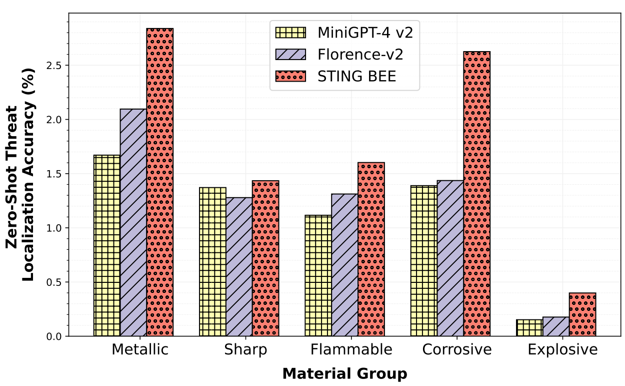

Figure 4. Comparison of zero-shot threat localization accuracy of STING BEE, Florence-v2, and MiniGPT-4 v2 across material groups in localizing objects with diverse material characteristics.

templates like `[grounding] Describe the baggage scan.` and `[refer] Please find <p>the threat category</p>`, with ground truth mapped to normalized bounding boxes for consistency. As shown in Fig. 4, STING-BEE outperformed MiniGPT-v2 [6] and Florence-2 [17], demonstrating robust performance across diverse material groups.

E. Additional Details on Data Augmentation

Existing approaches on data augmentation [8, 13] use single-direction projection methods limiting realism and orientation diversity. For the STCray dataset, our data augmentation approach is distinguished by a novel multi-view projection aspect enabled by our in-house CT scanner (see Sec. F.2). Using this scanner, we generated three CT scans of threat objects, rotating them in various orientations and projecting them into 2D X-ray images through a carefully designed technique. These X-ray images simulate the scanning of threat items from multiple perspectives. Subsequently, we integrated these scans into normal baggage scans using standard fusion and colourization techniques. The details of the approach are provided next.

Given $M(x, y, z)$ a 3D tensor representing a CT scan of a threat object, we apply a rotation $R_{\phi,\theta,\psi}(M)$, where ϕ, θ, ψ refer to the Euler angles representation. Afterwards, assuming an ideal narrow beam geometry and ignoring scatter, we approximate the projection of the threat item by computing the integration along the nominal z-axis[10]: $P(x, y) = \int R_{\phi,\theta,\psi}(M(x, y, z))dz$. The final image intensity I at any point is then calculated using the exponential attenuation model[10]:

$$I = I_0 e^{-\int R_{\phi,\theta,\psi}(M(x,y,z))dz} \tag{1}$$

STCray Caption:

"X-ray scan showing a power bank aligned horizontally, in the middle of a travel bag, with the power bank covered by the metal grid and hangers, along with some random cables for distraction, and there are other items like an umbrella and other normal passenger items."

Multi-Round Question-Answer Conversation:

- **Human:** Can you identify any potential security threats in this X-ray scan?
- **Assistant:** Yes, the scan contains a powerbank, which is a prohibited item.

- **Human:** Where is the powerbank located in the scan?
- **Assistant:** The powerbank is aligned horizontally in the middle of the travel bag.

- **Human:** What is obscuring the powerbank in the scan?
- **Assistant:** The powerbank is covered by a metal grid, hangers and various cables, which serve as distractions.

- **Human:** What color and shape should I look for to identify the powerbank?
- **Assistant:** You should look for a greenish rectangular block, as powerbanks appear with a greenish tinge due to their inorganic material.

- **Human:** Are there any other items in the scan that could be mistaken for a threat?
- **Assistant:** Yes, there are other normal items such as an umbrella, hangers and random cables, which could distract from identifying the powerbank's presence.

- **Human:** How do these normal items affect the detection of the powerbank?
- **Assistant:** The presence of normal items like cables and an umbrella can create confusion and distract from the identification of the powerbank.

- **Human:** What should be done about the powerbank in this scan?
- **Assistant:** The powerbank should be flagged for further inspection and potential removal from the baggage, as it is a security threat.

Figure 5. An instance of GPT-4 generated VQA instruction featuring the baggage scan and its corresponding caption from STCray at the top. The instruction-following data, generated using only the caption, is displayed below. Note that the baggage scan was not provided to GPT-4 during prompting and is included here solely for readability.

where I_0 is the initial beam intensity being exponentially attenuated by the projection path.

In the final stage, we fused a patch from a normal baggage scan with a threat image at a given location using pixel-wise multiplication. Both the baggage scan and the threat image are represented in grayscale. For colorization, we utilized a UNet architecture trained in a self-supervised manner to convert grayscale images into pseudo-colored scans. However, more advanced colorization techniques can be used, such as the recent method proposed by Duan et al. [8].

In summary, our method generates 2D projections with varying X-ray attenuation levels, introducing increased variability in the augmented scan data. This enhanced diversity effectively simulates real-world smuggling scenarios where threat items are positioned in unconventional ways within baggage to improve concealment. Furthermore, our approach enables automated annotation by implicitly generating threat masks within the augmented data.

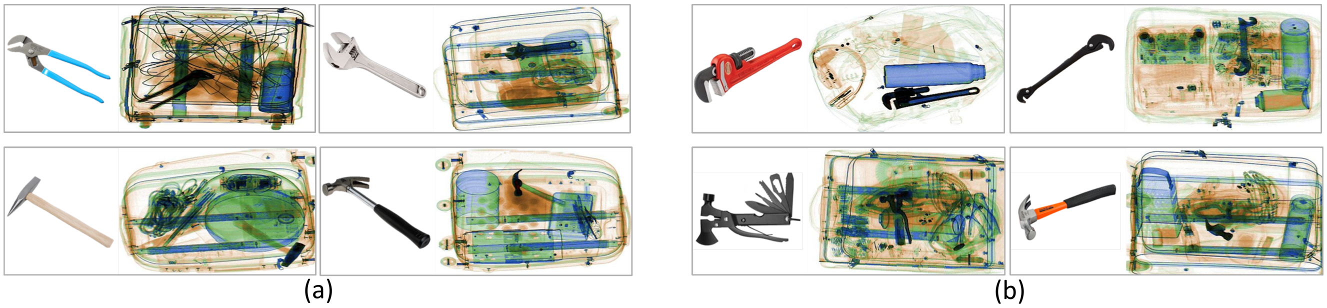

Figure 6. Illustration of train-test diversity and intra-category variability in the STCray dataset. The top row shows examples from the *Wrench* category: (a) *Pliers Wrench* and *Crescent Wrench* from the train set, and (b) *Pipe Wrench* and *Self-adjusting Wrench* from the test set. The bottom row highlights the *Hammer* category: (a) *Brick Hammer* and *Claw Hammer* from the train set, and (b) *Multi-tool Hammer* and *Framing Hammer* from the test set. A clear separation between train and test sets underscores the high intra-category variance, simulating real-world diversity, and challenging model generalization.

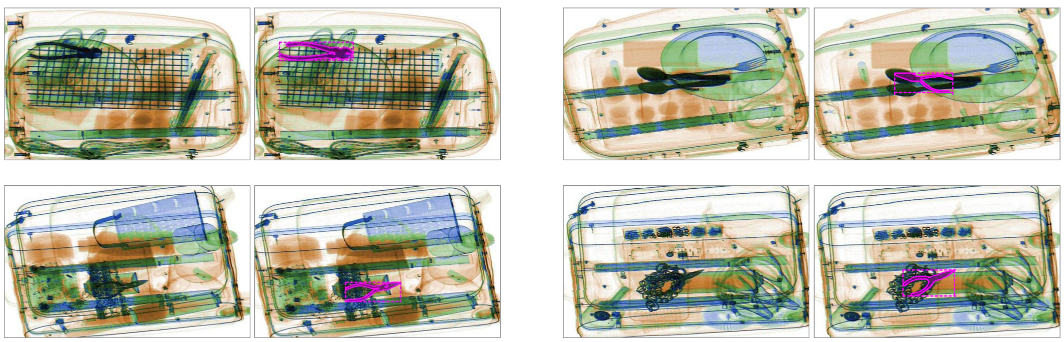

Figure 7. Illustration of diverse occlusion strategies in the STCray dataset across the train and test subsets. The training set (top row) employs techniques such as occluding the *Plier* with metallic grids (left) and disguising it with spoons (right). The testing set (bottom row) introduces novel strategies, including obscuring the *Plier* with a box of Integrated Chips (ICs) (left) and disguising it as an ornament using a chain (right).

Due to the limited size of our CT scanner's imaging chamber (as detailed in the equipment section Sec. F.2), we initially produced augmented scan data for the six threat items reported in Fig. 11 .

F. STCray Dataset Characteristics

The *Strategic Threat Concealment X-ray (STCray)* dataset introduces unique challenges that set it apart from existing X-ray security benchmarks, establishing itself as a pivotal resource for advancing research in baggage threat detection. Below, we provide a comprehensive overview of the characteristics and challenges that STCray brings to the research community.

- **Emerging Sophisticated Threats:**STCray addresses modern security challenges by incorporating a diverse range of contemporary threats, including 3D-printed firearms and improvised explosive devices (IEDs), alongside other prohibited items. Specifically, the dataset features three distinct 3D-printed gun designs: the single-shot *Liberator* [2], the minimalist and recent *Harlot* [1], and the pepper-box styled *Maverick* [3]. These designs were chosen to represent diverse shapes and structures, reflecting the variety encountered in real-world scenarios. Detecting 3D-printed firearms is particularly challenging due to their faint outlines and unconventional material properties, which blend with benign objects in X-ray scans (see Fig. 9).
Similarly, IEDs pose unique detection challenges due to their non-standard shapes and multi-component configurations. As depicted in Fig. 10, STCray incorporates both cohesive (compact) and dispersed IED types: cohesive designs consolidate explosive charges, detonators, and power sources into a single, unified threat, while dispersed designs distribute these components throughout the baggage, interconnecting them with wires. These variations further complicate detection, mirroring the diverse tactics employed in real-world scenarios.
- **Strategic Threat Concealment:** STCray is meticulously designed to reflect the complexities encountered in real-world baggage screening scenarios, using a carefully designed STING protocol , incorporating systematic threat concealment strategies that mimic smuggling tactics and concealment practices. Our collaborators from baggage screening units at the airport shared these smuggling prac-

Figure 8. The ANER K8065 X-ray Baggage Scanner used for the STCray dataset collection. On the right: the MARS Microlab 5X120 CT scanner which was employed for data augmentation Appendix E.

Figure 9. Left: *Maverick* 3D-printed firearm. Center: A sample baggage setup, concealing the firearm among everyday items. Right: The corresponding X-ray scan, illustrating the challenge of detecting faint outlines and identifying the firearm in a cluttered, realistic baggage scenario.

(a) (b)

Figure 10. Samples from the *Explosive* category: baggage preparation and the corresponding X-ray scan. (a) Cohesive (compact) IED with components grouped together, and (b) IED with components distributed across the baggage and connected by wires.

tices with us. Threat items are deliberately positioned and occluded by clutter and dense objects, with varying levels of overlap, material density, and angular placement. This systematic approach ensures that the dataset captures realistic concealment scenarios, pushing models to their limits. Fig. 12 reports examples of different concealment levels for three different threat items.

- **Diversity and Realism:** Unlike traditional datasets that often retain similar configurations across train and test sets—varying primarily in occlusion levels but preserving the same benign objects and threat items—STCray was designed to reflect the complexity and variability of real-world baggage scenarios. This diversity ensures realistic and challenging evaluation scenarios, mimicking the complexity and variability of real-world baggage screening (see Sec. F.1).

F.1. Intra-Category Variance

The train and test subsets of the STCray dataset both adhere to the STING protocol, ensuring systematic threat concealment and realistic variations. However, to introduce real-world relevance and challenge model generalizability, we have ensured intra-category diversity and diverse concealment strategies across the train and test subsets. For example, as shown in Fig. 6, the train set (top row) includes baggage scans featuring *Wrench* instances such as the *Pliers Wrench* and *Crescent Wrench* (left), while the test set contains entirely different *Wrench* types, such as the *Pipe Wrench* and *Self-adjusting Wrench* (right). Similarly, in the *Hammer* category (bottom row), the train set includes the *Brick Hammer* and *Claw Hammer* (left), while the test set introduces novel instances like the *Multi-tool Hammer* and *Framing Hammer* (right).

This diversity extends to the occlusion strategies employed across subsets. For instance, as shown in Fig. 7, the training set (top row) uses tactics such as metallic grids to occlude the *Plier* or spoons to disguise it. In contrast, the testing set (bottom row) introduces entirely new strategies, including obscuring the *Plier* with a box of Integrated Chips (ICs) or disguising it as an ornament using a chain.

These variations in both object structure and occlusion highlight STCray's emphasis on high intra-category variance and its ability to challenge models beyond the limitations of conventional datasets, where train and test splits often lack such diversity. By integrating these realistic complexities, STCray not only mimics real-world scenarios but also serves as a robust benchmark for evaluating model generalization and resilience to real-world variability, thereby addressing a critical gap in existing X-ray security datasets.

F.2. Scanner Details

The STCray dataset is collected using an ANER K8065 X-ray scanner Fig. 8. It is an advanced X-ray baggage scanner designed for security applications in airports, metro stations, and other security checkpoints. It features a tunnel size of 800 mm in width and 650 mm in height, accommodating a wide range of luggage sizes. The conveyor operates at an adjustable speed of 0.22 meters per second and can handle a maximum load of 200 kilograms. It is equipped with a high-resolution 17-inch LCD display and utilizes a dual-energy L-type photoelectric diode array detector with 12-bit depth, providing precise and detailed images. The scanner is equipped with software providing image process-

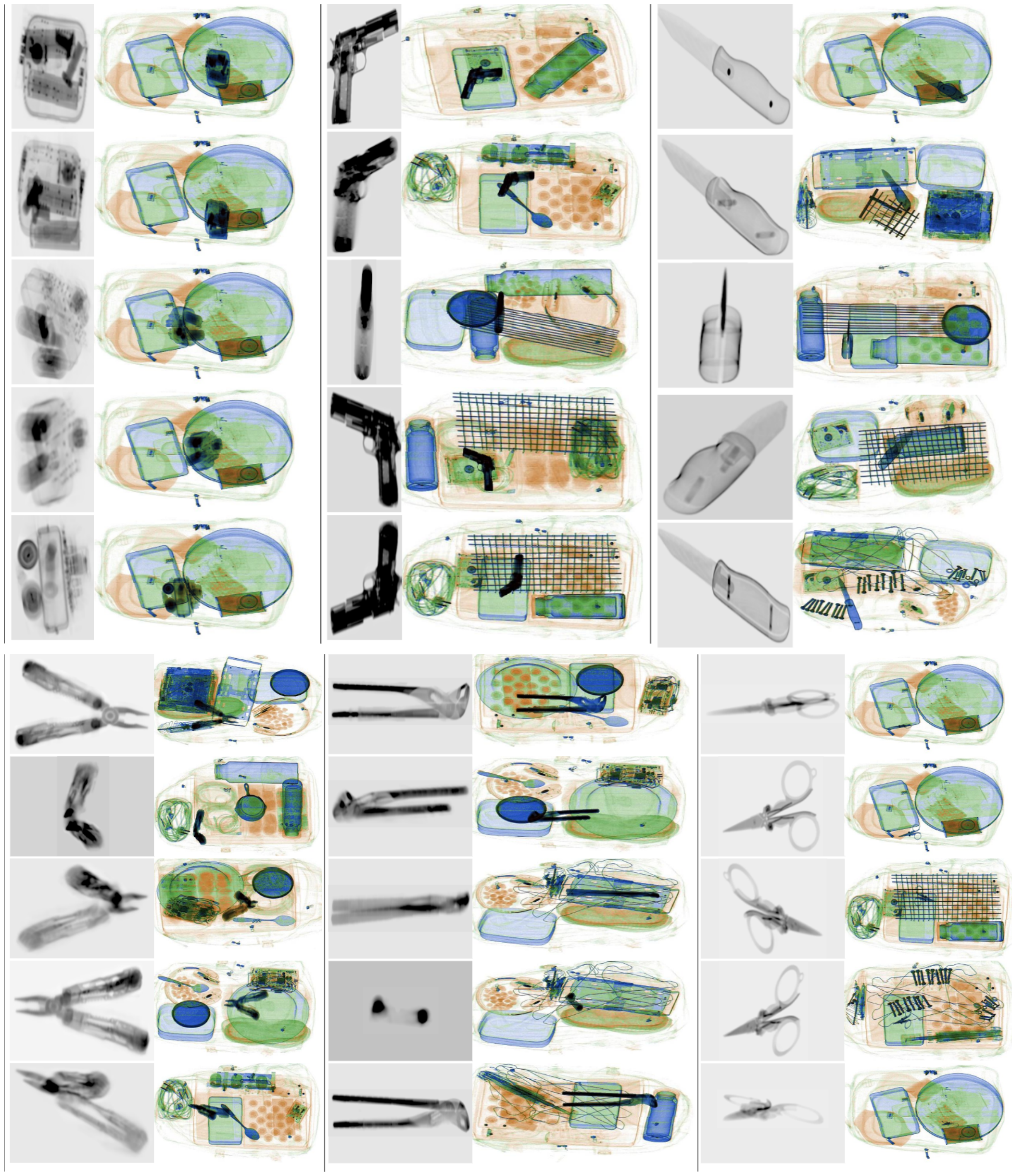

Figure 11. 2D X-ray projections, at different angles, from CT scans of threat items (explosives, guns, knives, pliers, wrenches, and scissors) and their augmented images. Notice the variation in the grayscale across the different projections reflecting different levels of X-ray attenuation. Note also that the atypical poses of threats generated through these projections create challenging instances, making threat detection more difficult. For example, the fourth and last instances of the wrench (3rd column) and the scissors (last column), respectively.

ing capabilities like edge enhancement, super image enhancement, high and low penetration display modes, and a magnifier function for partial enlargement. The machine boasts up to 40 mm steel penetration capability, effectively capturing concealed items within dense materials.

For the data augmentation, we used our in-house CT scanner model MARS Microlab 5X120 (Fig. 8), which offers a spatial resolution range of [50–200 μ mm]. Its chamber can accommodate objects with a maximum diameter of 100 mm and a length of 350 mm. Consequently, we were able to scan only a subset of the available threat items, including an explosive, a gun, a knife, pliers, a wrench, and scissors.

G. STING BEE: Additional Results

G.1. VQA Capabilities

To assess reasoning capabilities, we grouped the VQA categories by reasoning levels—Image-Wide Context (Instance

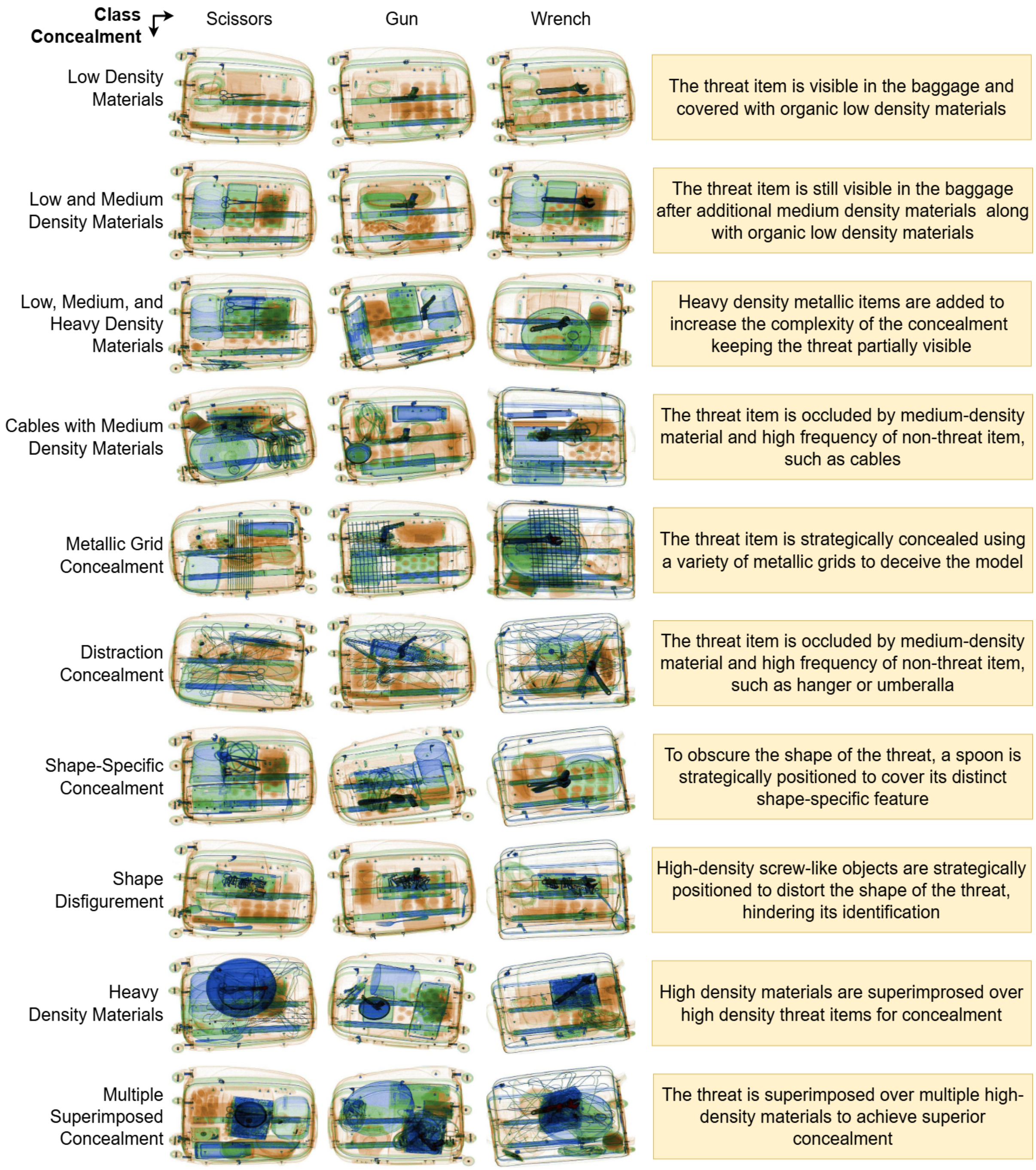

Figure 12. Illustrative examples of the 10 concealment levels utilized in the STCray dataset, demonstrating increasing complexity in obscuring threat items (scissors, gun, wrench) through diverse materials and strategies. The accompanying descriptions explain the concealment techniques and their effectiveness in reducing threat visibility.

Identity, Instance Counting, Instance Location), Object-Level Reasoning (Instance Attribute, Instance Interaction), and High-Level Domain Awareness (Complex Visual Reasoning)—and plotted model accuracy across these groups (see Fig. 13). Further analysis of the plot shows that STING BEE outperforms other models in both Image-Wide (65.64%) and domain-level (79.17%) tasks, demonstrating its comprehensive image-wide and domain-specific knowledge due to domain-targeted instruction tuning across multiple tasks.

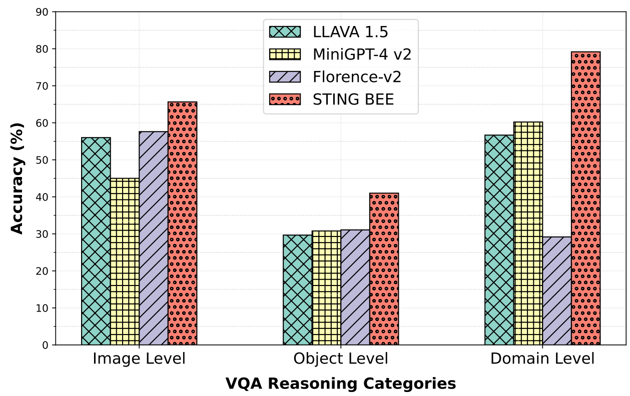

Figure 13. Model Accuracy Across Different Reasoning Categories in Visual Question Answering for X-ray Threat Detection.

Table 2. **Cross-Domain** comparison of different models on multi-class classification tasks. Models are trained on the STCray training set and tested on a combined SIXray and PIDray dataset, showing overall mAP and F1-score.

Model	mAP	F1 Score
EfficientNet-B4 [15]	9.5	9.0
GWFS [9]	5.3	6.6
DSACDIC [18]	3.7	6.7
ViT [7]	8.4	12.1
DeiT [16]	9.1	8.9
DINO [5]	11.4	8.3
CLIP [14]	8.1	3.3
LongCLIP [12]	9.5	4.7
STING BEE (Ours)	**30.1**	**36.2**

G.2. Cross-Domain Generalization

The results in Table 2 demonstrate the generalization ability of our proposed model, STING BEE, in multi-class classification tasks. To ensure a fair evaluation, all state-of-the-art and baseline models were trained on the STCray dataset and tested on a combined SIXray and PIDray dataset, with unique categories such as *Baton* and *Sprayer* removed to align the evaluation settings. Despite the competitive results from GWFS and DSACDIC, which are designed for cross-domain scenarios, on specific classes, STING BEE achieved the highest overall mAP of 30.1 and F1 score of 36.2, consistently outperforming all other models across most threat items, including Gun, Wrench, Hammer, and Powerbank. These results highlight STING BEE's ability to handle scanner-induced variability and intra-class differences, establishing it as a robust benchmark for cross-domain generalization in threat detection.

G.3. Comparison with General-purpose VLMs

Since STING-BEE is the first X-ray security visual AI assistant capable of scene comprehension, visual grounding, and VQA, we compared it with other general-purpose (open and closed-source) VLMs. The results in Table 3 show that

Table 3. Our STING-BEE outperforms Llama 3.2, Qwen2-VL, LLaVa (Finetuned) and GPT-4o. IL:Instance Location, CR:Complex Reasoning, IID:Instance Identity, IC:Instance Counting, M:Misleading, IA:Instance Attribute, II:Instance Interaction.

. Model	Scene Comprehension		Visual Question Answering (VQA)								Grounding	
	F1	mAP	IL	CR	IID	IC	M	IA	II	Overall	acc@0.5	acc@0.25
GPT-4o	19.4	18.1	21.5	16.2	36.3	29.6	25.2	19.02	18.3	31.3	-	-
LLaVa 1.5 (Finetuned)	28.3	22.2	35.2	59.4	77.8	36.1	15.9	48.2	25.3	45.1	-	-
Llama 3.2	13.4	17.3	26.5	29.3	36.3	20.8	11.6	25.4	13.3	23.5	1.9	5.3
Qwen2-VL	18.6	16.9	31.3	15.1	41.3	24.1	18.8	36.1	13.1	27.8	1.2	3.9
STING-BEE (limited captions)	22.1	18.2	25.3	41.3	69.2	35.4	20.5	42.2	27.6	41.8	5.3	14.1
STING-BEE	**34.7**	**29.8**	**49.2**	**79.2**	**80.0**	**45.2**	**27.8**	**52.8**	**35.0**	**52.8**	**8.7**	**21.5**

STING-BEE outperforms GPT-4o on both VQA and cross-domain scene comprehension tasks. STING-BEE also surpasses Qwen2-VL and Llama 3.2 in scene comprehension, VQA, and grounding capabilities, highlighting the need for domain-aware VLMs in X-ray security screening. To further analyze the relevance of the detailed captions in STCray, we trained STING-BEE with only image labels and box-level annotations. As shown in Table 3, STING-BEE with limited annotations still performs better than general purpose models; however, performs lower compared to STING-BEE trained with detailed annotations (Table 3).

G.4. Qualitative Results

Qualitative examples shown in Fig. 14 highlight STING-BEE's ability to excel across the key vision-language tasks, including scene comprehension, referring threat localization, visual grounding, and VQA, especially across diverse datasets, including STCray, SIXray, PIDray, and Compass XP, emphasizing its cross-domain generalization despite scanner variations. These results also highlight STING-BEE's robustness in handling high intra-categorical variability within threat categories, effectively addressing challenges like occlusion and object concealment in complex X-ray imagery.

To further illustrate these capabilities, Fig. 15 showcases STING-BEE's performance in scene comprehension, highlighting its ability to classify threat and non-threat items in complex baggage scans. Fig. 16 presents examples of visual grounding, demonstrating precise identification and localization of specific threat objects in the X-ray scans. Fig. 17 provides qualitative results for referring threat localization, emphasizing the model's ability to locate objects based on textual references. Finally, Fig. 18 showcases STING-BEE's effectiveness in handling diverse Visual Question Answering (VQA) tasks.

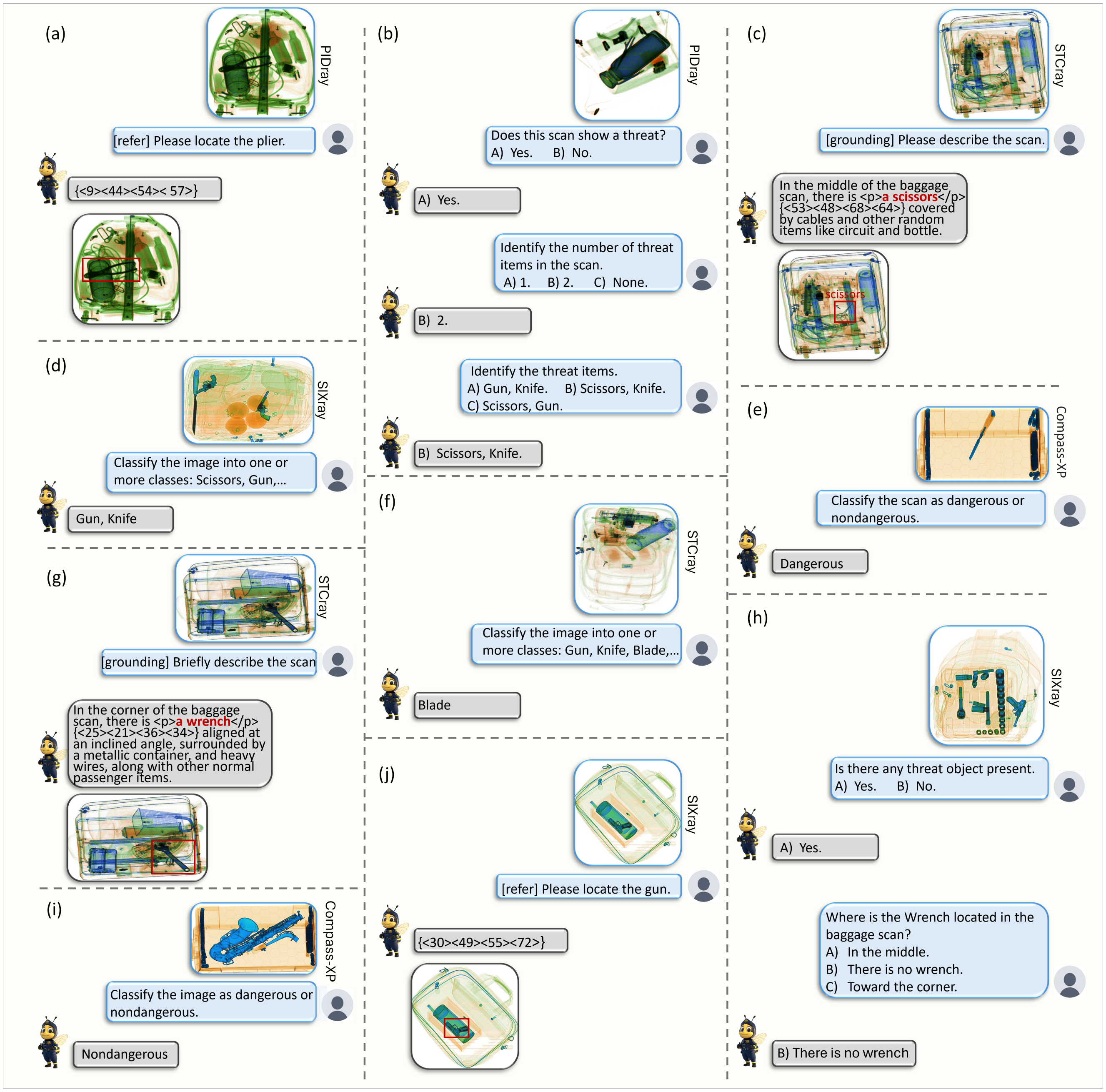

Figure 14. Qualitative examples showcasing the capabilities of STING-BEE across four vision-language tasks: Scene Comprehension (d, e, f, i), Referring Threat Localization (a, j), Visual Grounding (c, g), and Visual Question Answering (b, h). These examples span four X-ray security datasets— STCray, SIXray, PIDray, and Compass XP — illustrating STING-BEE's robustness and adaptability to diverse X-ray imagery.

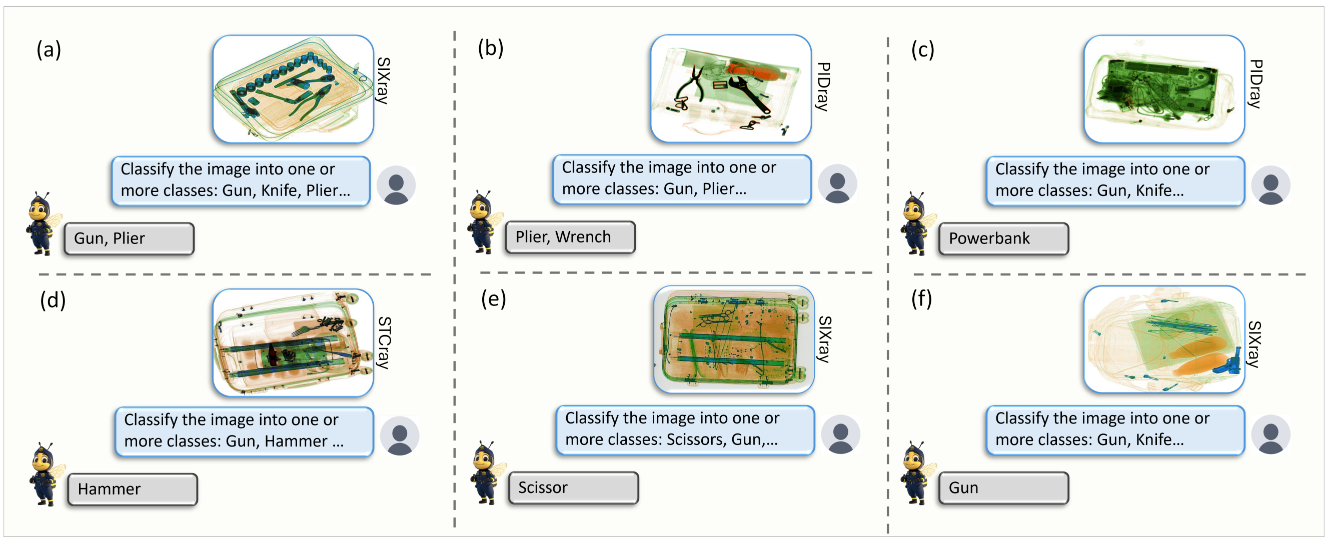

Figure 15. Scene comprehension qualitative examples showcasing the ability of STING-BEE to classify X-ray baggage scans into one or more threat-related classes. The images display diverse objects such as guns, pliers, wrenches, power banks, scissors, and hammers across different scenarios, highlighting the robustness of STING-BEE in understanding and categorizing threat items within X-ray imagery.

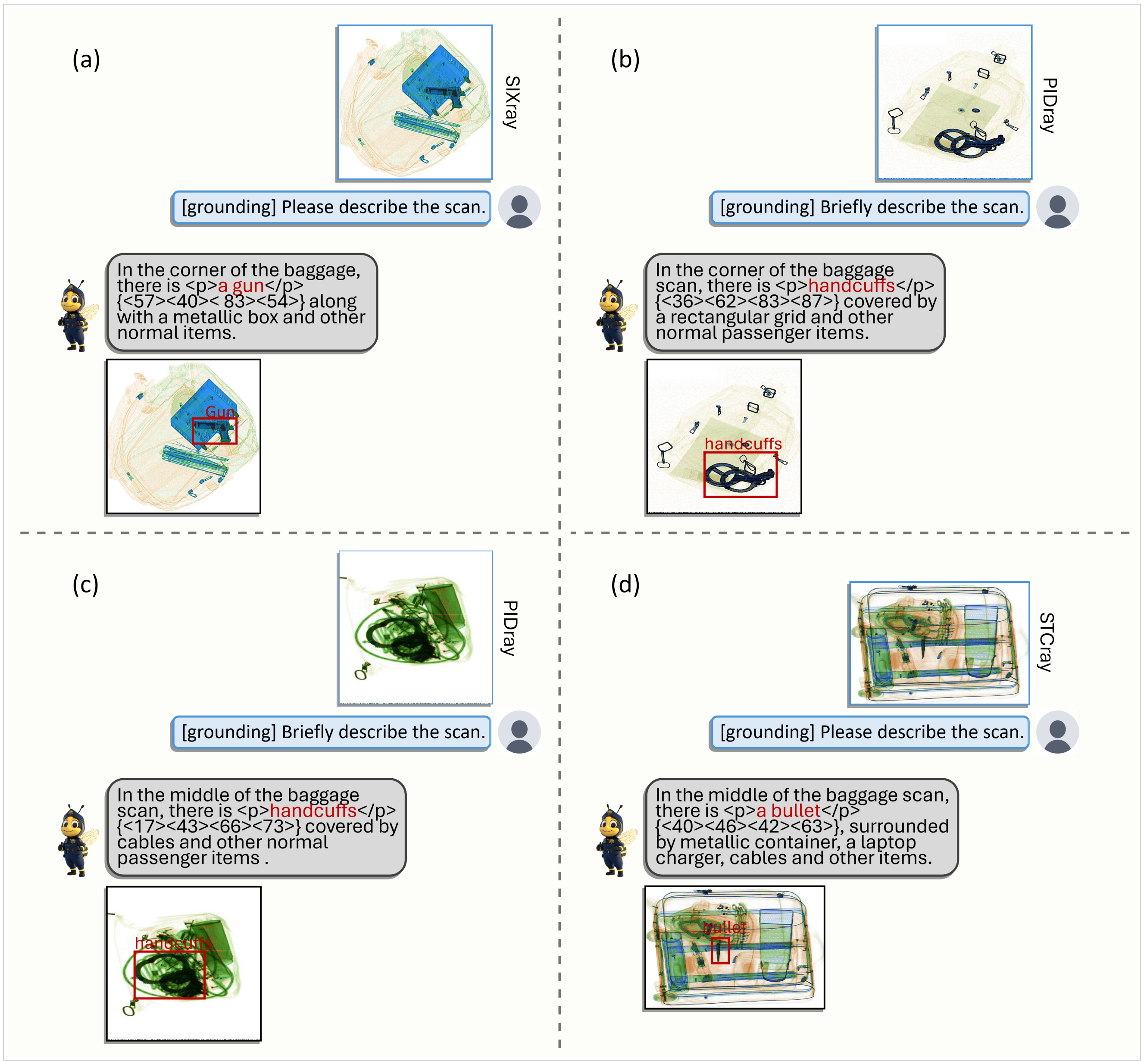

Figure 16. Visual grounding qualitative examples demonstrating STING-BEE's ability to describe and localize specific threat items in X-ray baggage scans. The system effectively identifies and highlights objects such as guns, handcuffs, and bullets within diverse scenarios.

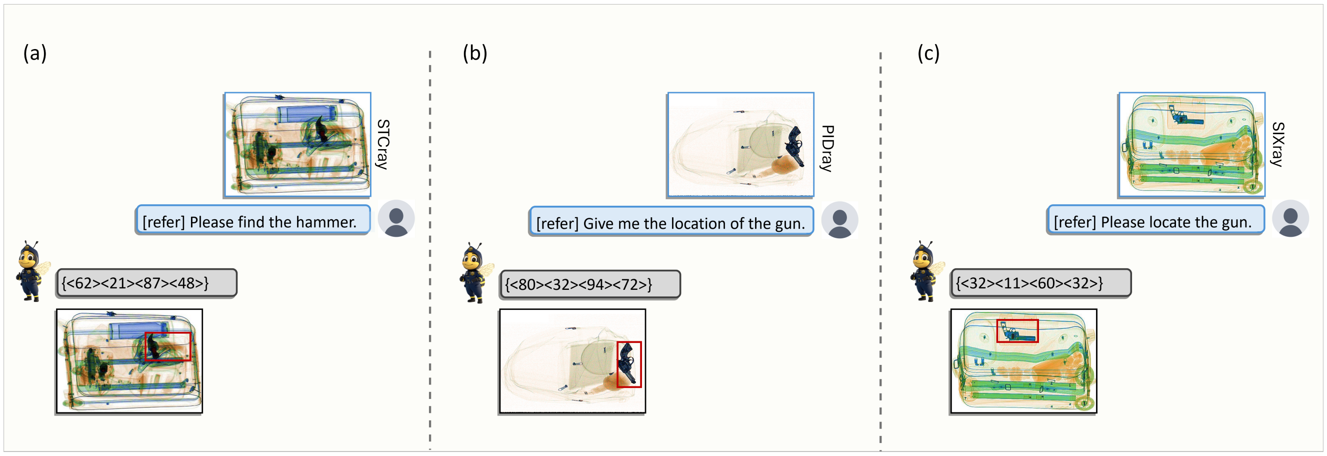

Figure 17. Referral threat localization examples showcasing STING-BEE's precision in identifying and localizing specific threat items in X-ray baggage scans. The system demonstrates its capability to locate different contraband objects, utilizing bounding box coordinates to highlight their positions within the scans.

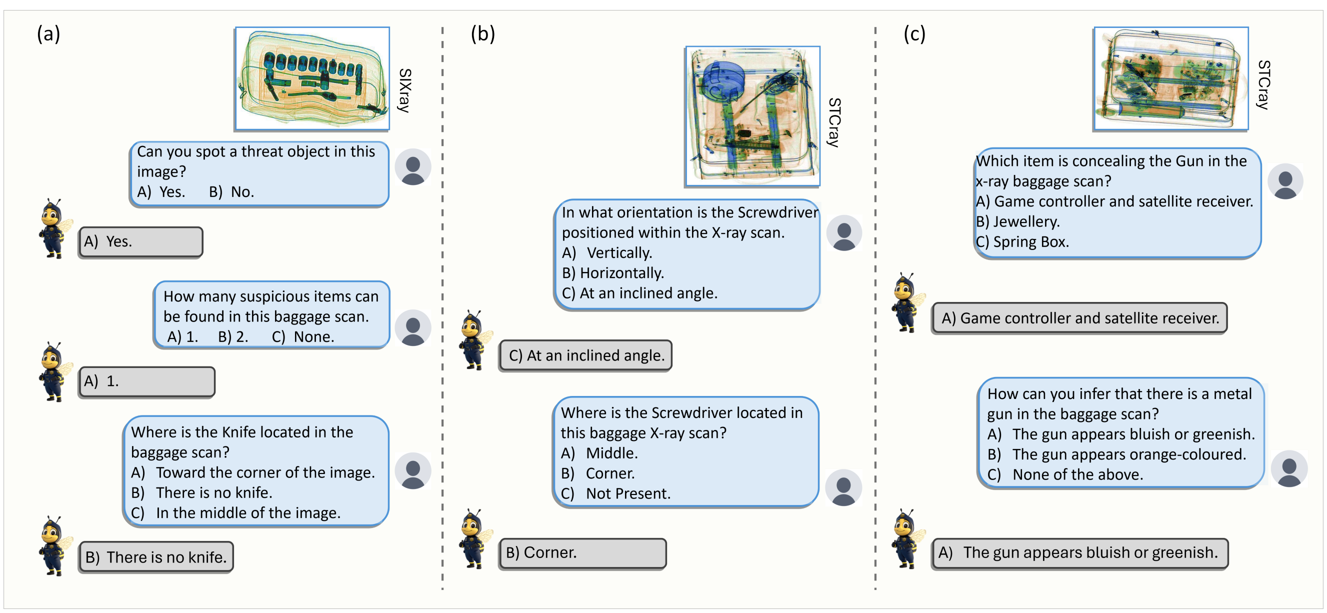

Figure 18. Qualitative examples showcasing the capabilities of STING-BEE in Visual Question Answering (VQA) across diverse question types: (a) Instance Identification, Instance Counting, and Misleading Question resolution, (b) Instance Location and Instance Attribute recognition, and (c) Instance Interaction and Complex Visual Reasoning.

References

[1] The harlot - print-a-22. https://www.print-a-22.com/the-harlot/. Accessed: 2024-11-17. 5

[2] Liberator (gun) - wikipedia. https://en.wikipedia.org/wiki/Liberator_(gun). Accessed: 2024-11-17. 5

[3] Pg22 - pilot hobbies. https://www.pilothobbies.com/pg22/. Accessed: 2024-11-17. 5

[4] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 9

[6] 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. 3

[7] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 9

[8] Luwen Duan, Min Wu, Lijian Mao, Jun Yin, Jianping Xiong, and Xi Li. Rwsc-fusion: Region-wise style-controlled fusion network for the prohibited x-ray security image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22398–22407, 2023. 3, 4

[9] Taimur Hassan, Samet Akçay, Mohammed Bennamoun, Salman Khan, and Naoufel Werghi. Unsupervised anomaly instance segmentation for baggage threat recognition. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–12, 2023. 9

[10] Hsieh John. Computed tomography: Principles, design, artifacts, and recent advances. In *SPIE Press.*, 2009. 3

[11] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024. 2

[12] Zhiyuan Li, Yichen Wang, Yutong Chen, Yuxuan Lu, Fan Yang, et al. Longclip: Introducing long-sequence transformers for text-image models. *arXiv preprint arXiv:2205.04735*, 2022. 9

[13] Domingo Mery and Aggelos K. Katsaggelos. A Logarithmic X-Ray Imaging Model for Baggage Inspection: Simulation and Object Detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, volume 2017-July, pages 251–259. IEEE Computer Society, aug 2017. 3

[14] 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. 9

[15] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6105–6114. PMLR, 2019. 9

[16] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, pages 10347–10357. PMLR, 2021. 9

[17] 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. 3

[18] Yewei Zhao, Hu Han, Shiguang Shan, and Xilin Chen. Deep subdomain alignment for cross-domain image classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2820–2829, 2024. 9