Supplementary Material for IRGPT: Understanding Infrared Modality with Cross-modal Curriculum on Real Infrared Image

Anonymous ICCV submission

Paper ID 361

A. Benchmark Building Details

A.1 Recognition, Scene and Security

The target category is extracted from labels and inserted as a key term into template sentences, while three distractor categories absent in the image are selected from the candidate set to form answer options. The distinctions between these three tasks manifest as follows:
a) The **Recognition** task exclusively uses images containing a single target category, which serves as the correct answer.
b) In the **Scene** task, each image is unambiguously associated with one scene category, emphasizing holistic image interpretation rather than individual object recognition.
c) Conversely, the **Security** task requires selecting non-existent targets within the image, making its correct answers complementary to those in the Recognition task.

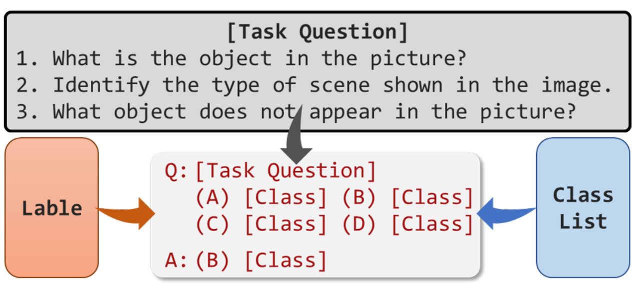


Figure 1: Recognition, Scene and Security task generation.

A.2 Relationship

The positional relationship task serves as a binary judgment task focused on verifying the accuracy of spatial relationship descriptions. Our methodology addresses two primary challenges in generating positional relationships: 1) formulating precise descriptions between target objects, and 2) accurately localizing reference objects.

For the first challenge, we consider six fundamental spatial relations: left, right, far, near, front, and back. Left/right relations are determined by comparing the horizontal positions of the centroids within predefined image regions, while front/back relations necessitate depth perception. To address this requirement, we integrate depth estimation through the MiDaS [?]. Near/far determinations combine 2D proximity in the image plane with depth proximity, applying empirically derived thresholds to classify relationships (far if exceeding threshold, near otherwise).

For the second problem, we adopt the same approach as Infrared-LLaVA [?], which involves left-to-right enumeration of same-category objects and designates the nth enumerated object as the reference target. The implementation framework is systematically illustrated in the accompanying figure.

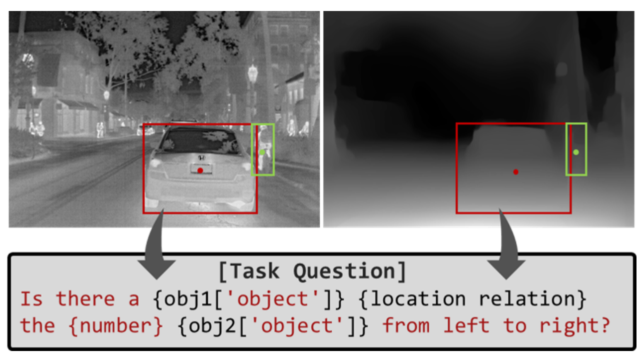


Figure 2: Relationship task generation.

A.3 Grounding

The grounding task primarily focuses on establishing correspondence between descriptions and target objects. Due to the limited appearance features in infrared images compared to visible images (where characteristics like color patterns can be effectively utilized), we propose two distinct approaches for target description, as illustrated in our framework. Firstly, we employ absolute positional descriptions based on image coordinates. Secondly, we implement relative positional descriptions through spatial relationships, which shares methodological similarities with relationship reasoning tasks. The former approach specifically utilizes a nine-grid division scheme for systematic region-based characterization.

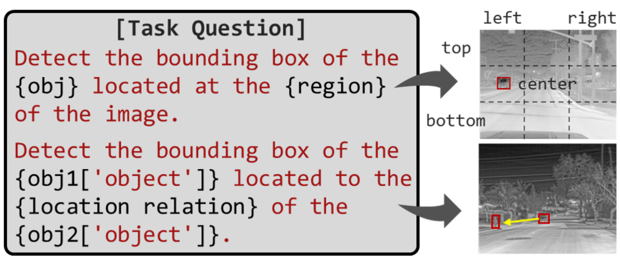


Figure 3: Grounding task generation.

A.4 Aerial Counting and Pedestrian Counting

Aerial Counting and Pedestrian Counting are two fundamental computer vision tasks focusing on estimating vehicle and pedestrian densities within images. The former operates on aerial perspectives captured by unmanned aerial vehicles (UAVs), while the latter analyzes surveillance camera viewpoints. Both methodologies have demonstrated significant practical utility in their respective operational domains, with Aerial Counting enabling large-scale traffic monitoring and Pedestrian Counting facilitating crowd management in urban surveillance systems.

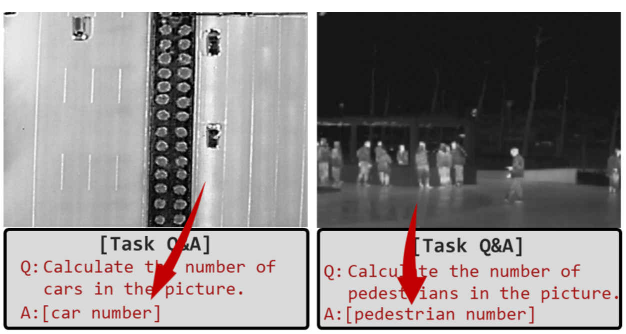


Figure 4: Aerial Counting and Pedestrian Counting task generation.

A.5 Location

The Location task is designed to determine the positional coordinates of targets within a specific category, constituting an advanced extension of the counting task. The spatial position of each target is operationally defined as the centroid coordinates of its corresponding bounding box (BBox). A schematic representation of this implementation framework is illustrated in the accompanying figure below.

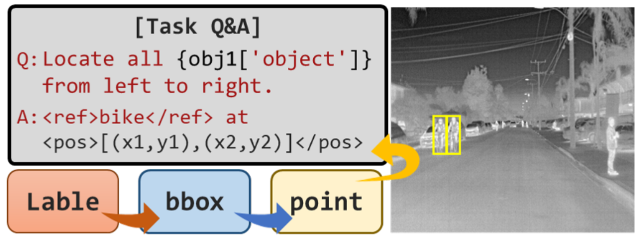


Figure 5: Location task generation.

A.6 Re-ID

The person re-identification (Re-ID) task represents one of the most prevalent applications in surveillance scenarios, aiming to verify whether individuals captured under different camera views share the same identity. In our experimental setup, we randomly sample eight pedestrian images and assemble them into a composite image, where only one instance matches the identity of the query image positioned at the top-left corner. This evaluation protocol essentially corresponds to measuring the Rank-1 accuracy rate, which serves as the primary performance metric. The detailed operational workflow is illustrated in Figure **??**.

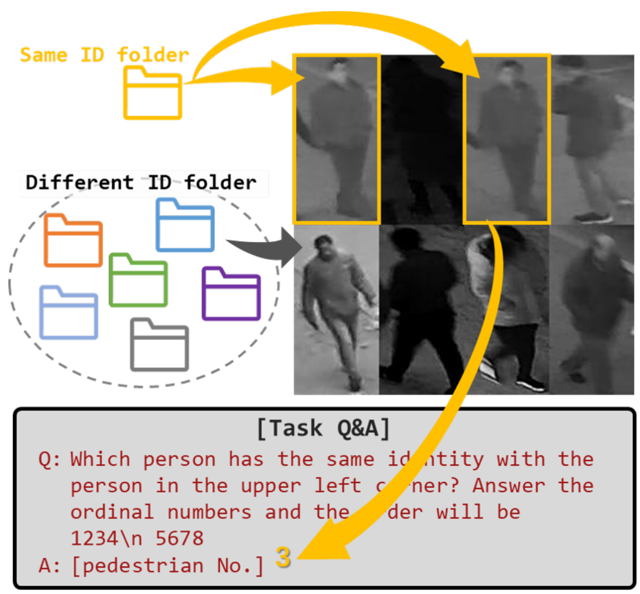


Figure 6: Re-ID task generation.

B. Caption after Low-light Enhancement

The low-light enhancement technique significantly alters the model's understanding of the scene by revealing previously obscured details and contextual relationships. In the original dim image, the model perceived a loosely organized parking lot with generic vehicle descriptions and no

The image depicts a nighttime scene of a parking lot, viewed from an elevated perspective. The primary focus is on several vehicles parked in the lot. The lighting in the image is dim, with the primary source of illumination coming from streetlights that cast a yellowish glow on the scene. The streetlights are positioned at various points around the parking lot, providing a limited but sufficient amount of light to outline the parked cars and the surrounding area.

There are at least five cars visible in the image. The cars are parked in a somewhat organized manner, with some parked closer to the streetlights and others further away. The cars appear to be of different models and sizes, but they are all parked in a manner that suggests they are not obstructing each other or the pathways within the parking lot.

The parking lot itself is relatively empty, with only a few cars present. The ground appears to be paved, and there are no visible markings or lines that would indicate parking spaces or lanes. The surrounding area is dark, indicating that the image was taken at night, and there are no visible buildings or structures in the immediate vicinity of the parking lot.

The image does not show any people, animals, or other objects besides the cars and streetlights. The overall atmosphere of the image is quiet and still, with the lack of movement or activity suggesting that the scene is captured during a time when the area is not busy.

In summary, the image shows a nighttime view of a parking lot with several cars parked in an organized manner. The primary light sources are streetlights that cast a yellowish glow on the scene, illuminating the parked cars and the surrounding area. The parking lot is relatively empty, and the surrounding area is dark, indicating that the image was taken at night.

The image depicts a nighttime scene of a street viewed from an elevated perspective, likely from a surveillance camera or a high vantage point. The street is illuminated by streetlights, casting a yellowish glow on the pavement and the parked vehicles. The lighting is sufficient to reveal the details of the scene but not bright enough to eliminate shadows, contributing to a somewhat dim and moody atmosphere.

In the foreground, there are several parked cars aligned along the street. The cars are mostly white or light-colored, and they are parked in a row, facing the same direction. The street appears to be relatively narrow, and the parked cars occupy a significant portion of the roadway, leaving only a narrow passage for vehicles to pass through.

On the left side of the image, there is a streetlight pole that stands out due to its height and the light it emits. The light from the streetlight casts a bright spot on the pavement and illuminates the area around it. The pole is positioned near the edge of the sidewalk, and its light extends across the street, providing visibility for the parked cars and the road.

In the background, there is a building that occupies a significant portion of the image. The building is dark, and its details are not clearly visible due to the low lighting. The building's facade is not distinctly visible, but it appears to be a multi-story structure, possibly residential or commercial.

There are no visible people or moving vehicles in the image. The scene is static, with the parked cars and the building providing the primary elements of the composition. The overall impression is one of a quiet, possibly residential area at night, with the streetlights providing the only significant source of illumination.

In summary, the image captures a nighttime street scene with parked cars, a streetlight, and a dark building. The lighting is provided by the streetlights, casting a yellowish glow on the scene. The composition is static, with no visible movement or people. The image conveys a sense of quiet and stillness typical of a residential area at night.

Figure 7: LLM outputs before and after low-light Enhancement.

background structures, constrained by limited visibility. After enhancement, the model identifies the scene as a narrow street with precise environmental cues: uniformly parked white vehicles, a multi-story building in the background, and directional lighting patterns from streetlamps. This transformation demonstrates how the enhancement amplifies subtle textures (e.g., car colors, architectural outlines) while preserving natural shadow dynamics, enabling the model to reconstruct spatial hierarchies (foreground vehicles vs. background structures) and recognize standardized urban layouts. Crucially, the technology balances brightness amplification with atmospheric authenticity – recovering actionable details like vehicle alignment for surveillance applications, yet retaining the scene's nocturnal character through controlled contrast rather than artificial over-illumination.

C. Training Loss

Here, we present the training loss trajectories with and without CL (Contrastive Learning), as illustrated in Figure **??**. The CL framework demonstrates effective training stabilization, achieving rapid convergence to a lower loss plateau during the initial optimization phase. Notably, the CL-based training exhibits minimal loss fluctuations, demonstrating a marked contrast between the two optimization strategies.

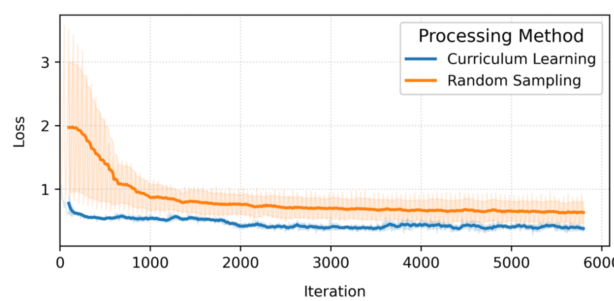


Figure 8: The curve of training loss.

ICCV
#361

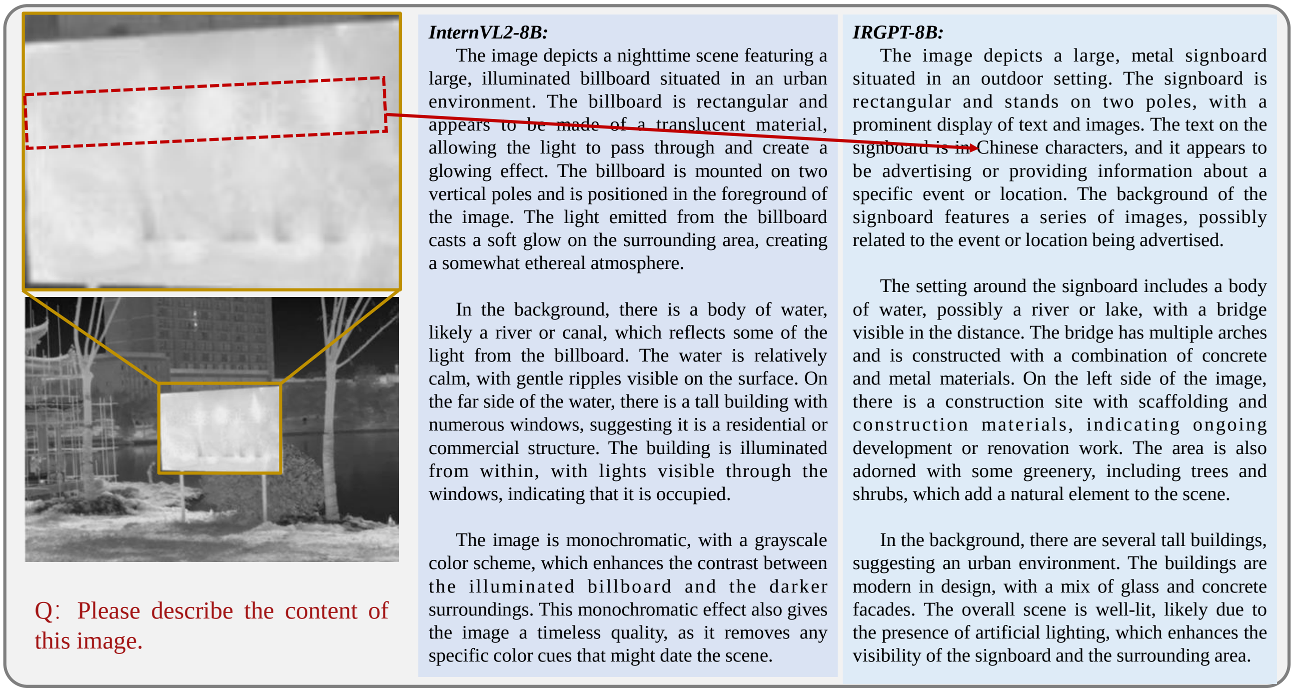


Figure 9: A comparative analysis of InternVL and IRGPT reveals that both models describe an identical image.

D. Comparative Analysis of InternVL-8B and IRGPT-8B

As shown in Figure **??**, the key distinction between the two descriptions lies in IRGPT-8B's superior ability to interpret infrared-specific details due to its specialized pre-training. While InternVL2-8B misidentifies the metallic signboard as "translucent" and focuses on atmospheric effects (e.g., "ethereal glow"), IRGPT-8B accurately recognizes structural elements like the metal-concrete bridge, construction scaffolding, and Chinese text on the signboard, reflecting its domain-specific understanding of material properties, contextual activity (e.g., ongoing construction), and semantic content. This precision, rooted in infrared training, enables IRGPT-8B to deliver actionable insights for real-world applications such as urban monitoring or infrastructure analysis, where material accuracy and functional context are critical.

E. Hyperparameter Sensitivity

As illustrated in Figure **??**, we conduct an ablation study on the hyperparameter M (number of hierarchical layers in the sampling strategy). The experimental results demonstrate that setting M=10 achieves peak performance while obtaining favorable values for both psum and nsum metrics.

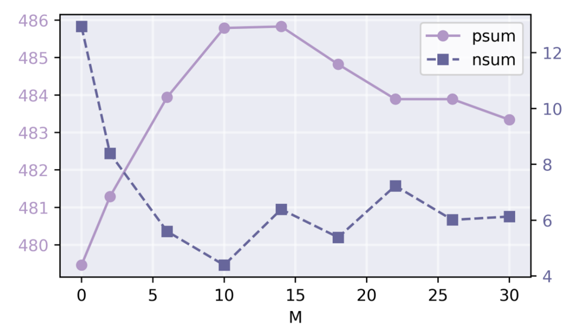


Figure 10: The curve of hyper-parameter ablation.

F. Dataset Ratio of LLM-base and Rule-based

As depicted in Figure **??**, we present the proportionate allocation between LLM-based and rule-based data sources.

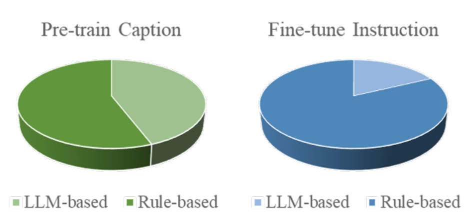


Figure 11: Dataset Ratio of LLM-base and Rule-based.

Question:
Please give a caption in detail.

Answer:
The image is an aerial view of a street scene. The street is divided into two lanes by a double yellow line, indicating that it is a two-way road. On the left side of the street, there is a sidewalk with a row of trees and a small garden area. The garden area has several plants and a small structure, possibly a shed or a small building.

On the right side of the street, there is a parking area with several cars parked in various spots. The parking area is adjacent to a larger open space that appears to be a construction site or a vacant lot. There are some construction materials and equipment visible in this area.

In the middle of the street, there are a few vehicles, including a car driving in the left lane and a few cars parked on the right side of the road. There are also a few pedestrians walking on the sidewalks. The overall scene suggests a quiet, possibly suburban or rural area with some ongoing construction or development.

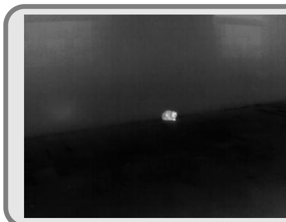
Question:
Please give a caption in detail.

Answer:
The image depicts a nighttime street scene with several elements. In the foreground, there are two cars parked on the side of the road. The car on the left is a white sedan, and the car on the right is a silver sedan. Between these two cars, there is a bicycle parked against a tree.

In the background, there are several people walking on the sidewalk. They appear to be engaged in conversation. Behind them, there are more trees and a building with a bright light shining from a window. The street is lined with trees, and there are power lines running across the image. The overall scene suggests a quiet urban area at night.

Question:
What is the shape of the top of the valve?

Answer:
The top of the valve is circular.



Question:
What is the cat doing in the image?

Answer:
The cat is sitting on the ground, facing towards the right side of the image.

Question:
What can be inferred about the time of day in the image?

Answer:
The image depicts a nighttime street scene.

Question:
What time of day does the image likely depict?

Answer:
The image likely depicts daytime, as indicated by the shadows cast by the trees and street furniture.

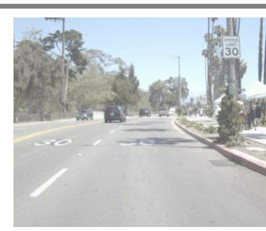


Figure 12: Some dialogue examples.

G. Model Q/A Presentation

As demonstrated in the Figure **??**, we present several dialogue examples of IRGPT, highlighting its robust image comprehension capabilities. The model accurately perceives small targets and subtle image details, exemplified by the cat's posture in the fourth example. Notably, IRGPT infers whether infrared images were captured during daytime or nighttime based on subtle cues, a task even challenging for human observers.

H. Source Data Directory

Table 1: Source datasets in details.

Index	Dataset name	Data size
1	LLVIP	15.5K
2	KAIST	95.3K
3	M3FD	8.4K
4	VEDAI	1.2K
6	LSOTB-TIR	643K
7	RGBT234	234K
8	LasHeR	734.8K
9	VTUAV	1.7M
10	MFNet	1.6K
11	IRSTD-1k	1K
12	SIRST-AUG	8.5K
13	BU-TIV	60K
14	VisDrone-DroneVehicle	28.4K
16	MassMIND	2.9K
17	ISDD	1.3K
18	CDFAG	88K
19	WideIRSTD	9K
20	DMIST	9K
21	JU-VNT	2.6K
22	POP	7.8K
23	RGBTCrowdCounting	2K
24	VT5000	5K
27	Industry	0.4K
28	VT1000	1K
29	VT821	821
30	UVT2000	2K
31	FLIR	10K
32-50	Web Data(several)	66K
51	RGBNIR	1K
52	UAV_RGB-T_2400	2.4K
53	LLCM	65.8K
54	RegDB	9K
55	SYSU-MM01	47K
56	ThermalWorld	8.1K
57	solar_cell_EL_image	36.5K
58	sirst_aug	18.1K
59	tirsequences	30K
60	DroneRGBT	7K
61	UVT20K	20K
62	time-sensitive	21.8K
63	FLIR_ADAS	66.8K

I. Extra Exploration

Table 2: Extra experiment of zero-shot case.

Case	psum	nsum
Pre-train on RGBT pairs (RGB)	310.54	102.69
Pre-train on RGBT pairs (T)	319.39	98.40
IRGPT-8B (main result in paper)	328.65	93.87
IRGPT-26B	332.68	87.36

We supplement several special cases in Table ??.

First, for the subset of IR-TD data with corresponding visible-light images, we exclusively use their RGB versions for training and evaluation. Comparative experiments using only IR counterparts reveal that RGB-based training still contributes to performance improvement, potentially due to the closer distribution alignment between RGB and T-version images. However, T-version training outperforms RGB-based approaches while remaining inferior to full-scale pre-training results.

Furthermore, we investigate larger-scale models and observe significant improvements in evaluation metrics.

J. Unaligned datasets

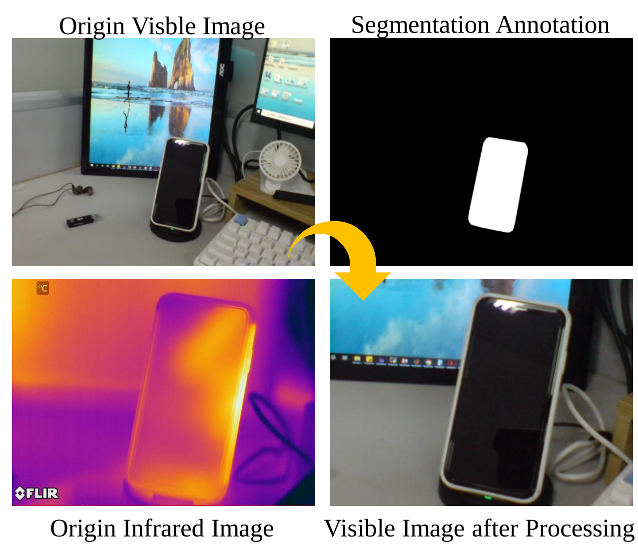


Figure 13: The processing of unaligned datasets.

We present our approach for handling unaligned RGB-T image pairs, which primarily leverages target annotations as anchor points to crop the wider field-of-view visible image, thereby ensuring semantic consistency within the image pair.

References

[1] Shixin Jiang, Zerui Chen, Jiafeng Liang, Yanyan Zhao, Ming Liu, and Bing Qin. Infrared-llava: Enhancing understanding of infrared images in multi-modal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8573–8591, 2024.

[2] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022.