Infrared-LLaVA: Enhancing Understanding of Infrared Images in Multi-Modal Large Language Models

Shixin Jiang[1], Zerui Chen[1], Jiafeng Liang[1], Yanyan Zhao[1], Ming Liu[1,2*], Bing Qin[1,2]
[1]Harbin Institute of Technology, Harbin, China
[2]Peng Cheng Laboratory, Shenzhen, China
{sxjiang, zrchen, jfliang, yyzhao, mliu*, qinb}@ir.hit.edu.cn

Abstract

Expanding the understanding capabilities of multi-modal large language models (MLLMs) for infrared modality is a challenge due to the single-modality nature and limited amount of training data. Existing methods typically construct a uniform embedding space for cross-modal alignment and leverage abundant visual image data to understand infrared images indirectly. However, they ignore the supervisory signals of infrared-modality-specific attributes, which may lead to a biased understanding of infrared images. To address this issue, we propose a debating multi-agent generation system that transfers knowledge from visible images to generate infrared image-text pairs and infrared instruction data. Moreover, we construct an infrared question-answering benchmark based on common infrared tasks. Experimental results from incremental fine-tuning on existing models and our Infrared-LLaVA-7B trained from scratch on infrared data demonstrate the effectiveness of the generated data and the feasibility of the generation approach.

1 Introduction

With the rapid development of large language models (LLMs) (OpenAI, 2024; Touvron et al., 2023; Chiang et al., 2023), existing multi-modal large language models (MLLMs) have shown impressive capabilities on generic visual modality tasks such as image-text retrieval and phrase grounding (Dai et al., 2024; Lin et al., 2023; Liu et al., 2024; Li et al., 2023a; Bai et al., 2023; Peng et al., 2023). But less attent ion has been paid to infrared images one of the scarce modalities (Su et al., 2023; Han et al., 2023), due to single-modality nature and low resourc of training data.

The infrared images have more stable information compared to visual images which have large information loss in low light or other extremely harsh environments (Jia et al., 2021; Liu et al., 2020). Thus it is worth exploring and researching how to utilize the limited data sources of the infrared images to align the infrared images' feature space with the LLM's text embedding space, and to further perform infrared instruction supervised fine-tuning. This would enable a MLLM to correctly accept infrared image inputs and perform infrared image question-answering tasks.

Previous work (Su et al., 2023; Han et al., 2023) has achieved some infrared image comprehension using only visible image-text pairs and visible image instructions for alignment and instruction fine-tuning, based on the unified embedding space implemented by ImageBind (Girdhar et al., 2023). However, the method does not involve the infrared image encoder during alignment and instruction fine-tuning. The lack of supervisory signals for infrared images may lead to biased understanding of infrared images. Hence it is necessary to construct infrared image-text pairs and infrared instructions to enhance the model's understanding ability for infrared images. Meanwhile, given to the lack of benchmarks for evaluating infrared-related capabilities of MLLMs, it is necessary to establish an infrared multi-modal question-answering benchmark to assess MLLM's understanding ability for infrared images.

To address these issues, we propose a debating multi-agent system. The system generates infrared image-text pairs and infrared instructions by transferring knowledge from visual image-text pairs and corresponding bounding box annotations, as illustrated in Figure 1. The system utilizes multiple independent GPT-3.5 models to construct multiple agents. Each agent respectively performs infrared image-text pairs generation, instruction generation, and instruction evaluation. Furthermore, the system ensures the quality of generated data by self-updating through the debates between different agents. Most importantly, the entire process is auto-

*Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 8573–8591

mated and does not require training, saving a significant amount of time and economic costs compared to annotating the data from scratch. Additionally, we build an infrared template question-answering benchmark based on annotated real infrared images to evaluate the MLLM's understanding ability for infrared images, as shown in Figure 2. This benchmark is constructed based on common infrared task datasets from multiple infrared scenarios, with high consistency and accuracy.

We use this multi-agent to generate 118k synthetic infrared images, 500k synthetic infrared image-text pairs, and 12k synthetic infrared instruction data based on COCO (Lin et al., 2014), called Infrared Instruction Dataset. We demonstrated the effectiveness of the synthetic instruction data through incremental fine-tuning on other models, as show in Figure 5. Specifically, with the aid of this dataset, we developed an infrared multi-modal large language model called Infrared-LLaVA-7B, as show in Figure 3. Infrared-LLaVA-7B outperformed Pandagpt-7B, Pandagpt-13B and imagebind-LLM trained in the infrared template benchmark, showing its competitive performance. Our contributions are as follows:

- We build a debating multi-agent system and generate infrared image-text pairs and instructions based on the visual image dataset. This provides a cost-friendly approach for constructing instruction data for scarce modalities lacking paired text and instructions.
- We construct the first infrared question-answering benchmark based on existing infrared image datasets, aiming to evaluate models' understanding ability for infrared images in typical scenarios, to the best of our knowledge.
- We validate the feasibility of indirectly aligning the infrared image embedding space and the text embedding space of LLMs using visible images by comparing the actual performances of indirect alignment and direct alignment.
- We build the first specific infrared multi-modal large language model, Infrared-LLaVA-7B. We utilize generated infrared data for both alignment and supervised fine-tuning stages.

2 Related Work

Multi-modal Large Models. In recent years, researchers have introduced visual multi-modal large-scale models, which integrate the large language models with pretrained visual encoders (Li et al., 2023a; Dai et al., 2024; Liu et al., 2024; Bai et al., 2023; Zhang et al., 2023). Most MLLMs follow a two-stage paradigm, utilizing large-scale image-text pairs for modality alignment in the pre-training stage and multimodal instruction data for supervised fine-tuning stage. Our work also adopts this paradigm as our training pipeline.

Scarce-Modality Large Multi-Modal Models. Compared to MLLMs that focus on modalities such as video and images (Li et al., 2023b; Bai et al., 2023; Liu et al., 2024), some projects like ImagebindLLM and Pandagpt (Han et al., 2023; Su et al., 2023; Zhang et al., 2023) utilize the same paradigm to understand modalities with scarce data, including infrared and depth images. Building upon unified embedding spaces achieved by cross-modal alignment encoders such as ImageBind (Girdhar et al., 2023), they leverage existing visible image-text pairs and instructions for training to understand infrared images. In our work, we further enhance the understanding of infrared images by conducting two-stage training using generated infrared-text pairs and infrared instructions. Compared to training on visible images data, we achieve better infrared image understanding performance with our generated infrared data.

Multi-Agent Systems. In recent years, the rise of large language models (LLMs) has paved the way for new directions in research on multi-agent systems. In multi-agent systems, collaboration among multiple agents has been widely applied to practical problems such as literary translation (Wu et al., 2024) and confidence verification (Yang et al., 2024). However, with increasing demands for system robustness (Zhu et al., 2024) and complex problem reasoning (Liang et al., 2023), introducing debating mechanisms has emerged as a new approach to enhance the performance of multi-agent systems. In the field of text generation, researchers have also begun to explore the introduction of debating mechanisms in multi-agent systems to improve the quality of generated texts (Kim et al., 2024; Chan et al., 2023). Therefore, we aim to enhance the quality of synthesized captions by incorporating a debating mechanism into the multi-agent system, and further explore text generation methods based on debating.

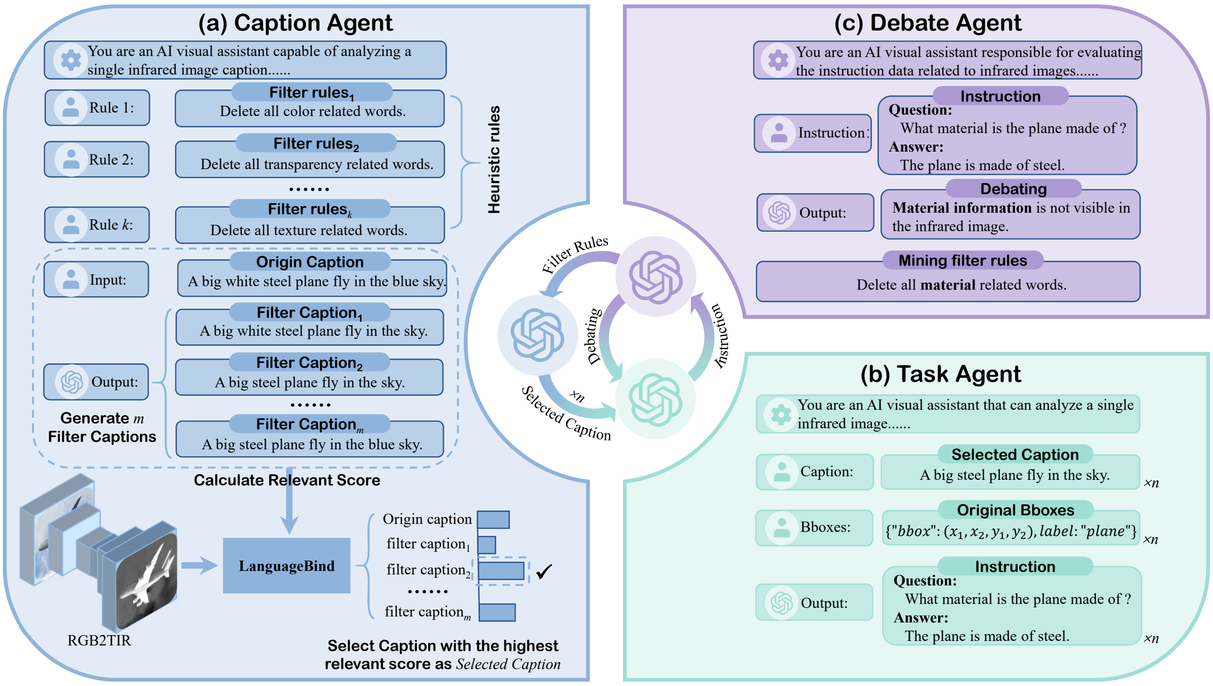


Figure 1: **The Multi-Agent System.** The multi-agent system includes three agents: Caption Generation Agent, Debate Agent, and Task Generation Agent. In addition to independently performing their own functions, these agents also interact with each other.

3 Infrared Instruction Dataset

In this section, we present our debating multi-agent system to build the Infrared Instruction Dataset. Subsection 3.1, Subsection 3.2, and Subsection 3.3 explain the composition and workflow of each agent. Subsection 3.4 describes the process of interaction in the multi-agent system. As a result, we generate 118k synthetic infrared images based on COCO(Lin et al., 2014), 500k paired captions, and 12k infrared image instructions.

3.1 Caption Agent

The Caption Agent is composed of three key models: the image translation model sRGB-TIR (Lee et al., 2023), the large language model GPT-3.5 (OpenAI, 2024), and the multimodal alignment model LanguageBind (Zhu et al., 2023a), as shown in Figure 1(a). The Caption Agent transforms the visible image-text pairs into the infrared image-text pairs. Then the Caption Agent outputs multiple filtered captions to the Task Agent and updates its filtering rules when the The Debate Agent provides the mining filter rules.

Infrared images translation. We utilize the image translation model sRGB-TIR (Lee et al., 2023) to translate the visible images into infrared images. This approach preserves prior knowledge from real visible images and ensure the consistency of scene structure and object authenticity in the translated infrared images. Additionally, it maintains high semantic relevance between the translated infrared images and the original captions paired with visible images (Gao et al., 2023).

Filtering based on heuristic rules. The original caption paired with visible images describe features like color and transparency, which are not discernible in infrared images (Gao et al., 2023). We first manually design several basic heuristic filtering rules (Hu et al., 2014) and then these rules will be automatically update using the output of the Debate Agent. Prompted by the filtering rules, the GPT-3.5 acts as an infrared filter to remove words from the captions of visible images that are not typically used to describe infrared images, while preserving most original semantic information.

Filtering based on image-text matching. We set the temperature of GPT-3.5 to 0.8 and generate 5 filtered captions for each visible image-text pair. Utilizing LanguageBind's Text Encoder and Infrared Encoder (Zhu et al., 2023a), we compute the similarity scores between these 6 captions (including filtered five and original one) and the translated infrared image. The caption with the highest score is selected as the final caption. By this way, the selected caption can maintain a high semantic relevance to the translated infrared image.

3.2 Task Agent

The Task Agent is comprised of an independent instance of GPT-3.5. This agent amalgamates original bounding box annotations with multiple filtered captions from the Caption Agent to generate instructions, which is then transmitted to the Debate Agent, as illustrated in Figure 1(b). Based on the debating result returned by the Debate Agent, the Task Agent regenerates the instruction accordingly.

Insturction generation. For each generated infrared image, corresponding filtered captions and bounding boxes are combined and then fed into GPT-3.5. In this way, GPT-3.5 can "see" the generated infrared image from a comprehensive perspective (Liu et al., 2024). And different crafted instructions and few-shot learning enable GPT-3.5 to generate three types of instruction tasks: complex reasoning, multi-round dialogue, and detailed description (see examples in Appendix B).

Insturction regeneration. The Task Agent conveys the generated instruction to the Debate Agent for evaluation. If the Debate Agent determines that the generated instruction do not meet the requirements, the Task Agent will regenerate a instruction with the failure reason. The process will continue until the Debate Agent returns "Accepted" or reaches the maximum times of repetitions. If "Accepted", the generated instruction will be output to the environment. Otherwise, "Generation failed" will be output.

3.3 Debate Agent

The Debate Agent is another independent instance of GPT-3.5. This agent evaluates whether the instruction generated by the Task Agent meets with requirements. The Debate Agent conveys the debating result to the Task Agent. Additionally, it extracts filtering rules from the debating result, which are then sent to the Caption Agent to update its filtering rules. As shown in Figure 1(c).

Instruction debating. The Debate Agent evaluates whether the generated instructions align with the characteristics of the infrared images based on relevant prompts and its own knowledge. If the generated instructions meet the characteristics, the Debate Agent returns "Accepted" as the debating result. Otherwise, the Debate Agent outputs the reasons for non-compliance as the debating result. The debating result is then sent to the Task Agent to regenerate the instructions.

Mining filtering rules. If the debating result is not "Accepted," the Debate Agent will extract filtering rules from the debating result. For example, if the debating result is "Material information is not visible in the infrared image" the extracted filtering rule would be "Delete all material-related words". This rule is then sent to the Caption Agent to update its filtering rules.

3.4 Working process

Initialization and external environment setup. First, we establish the basic filtering rules for the Caption Agent and define the prompt for the GPT-3.5 within each Agent. At each execution, the external environment feeds a COCO image and its corresponding multiple captions and bounding box annotations. The next execution starts after receiving the output from the Task Agent.

Agent Operation and Interaction. During each execution, the Caption Agent first translates the input visible image into an infrared image and sequentially filter the multiple captions paired with the visible image. Then the Task Agent generates instruction based on the filtered captions and the bounding boxes. The Debate Agent evaluates the instructions generated by the Task Agent and extracts filtered rules to the Caption Agent. Finally, the Task Agent outputs the result, while the Caption Agent updates its filtering rules. The entire interaction process is illustrated in Figure 1.

4 Infrared Template Benchmark

Existing research lacks a benchmark for question-answering tasks specifically for infrared images. To further evaluate the MLLM's ability to understand and respond to infrared images and commands, we build an infrared benchmark based on existing infrared dataset, as shown in Figure 2.

4.1 Data Collection

Collect real infrared datasets. We collect a total of 62 datasets including LLVIP and FLIR from existing research works. More analyses of the collected data is shown in Appendix A. We design an evaluation function for infrared dataset based on imaging quality and diversity, as shown in Equation (1).

$$D_{score} = 0.4 \cdot Q_{score} + 0.6 \cdot N_{score} \quad (1)$$

Where Q_{score} is the average score of the IQA model Q-Alion (Wu et al., 2023) on one hundred

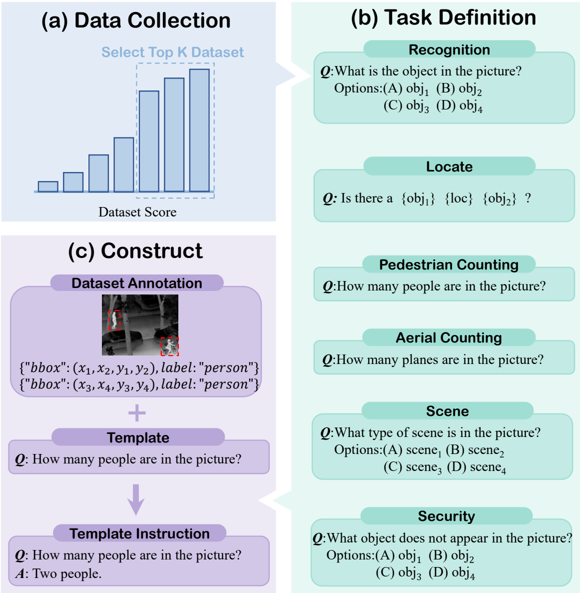


Figure 2: **Template-Based Construction.** The process includes three steps: filtering the dataset, tasks defining and combining annotations with templates.

sampled images of each dataset, and N_{score} is the number of object categories annotated in each dataset. Considering the scores and the richness of tasks, we retain six datasets, including tasks such as pedestrian recognition (Jia et al., 2021), object tracking (Liu et al., 2020), object detection (Sun et al., 2022), and scene classification (Brown and Süsstrunk, 2011).

4.2 Task Definition

The tasks in this benchmark are derived from the common infrared imaging tasks. Therefore, the tasks in this benchmark cover different levels of semantic information in infrared images, enabling a comprehensive assessment of the model's ability to perceive and understand infrared images.

Recognition and Scene. The recognition task and the scene task evaluate the ability of the model to recognize objects and scenes in infrared images. In both tasks, the model have to select the option that best matches the image from four candidates.

Counting. The aerial counting and pedestrian counting tasks evaluate the model's ability to count objects of the same category in different infrared scenarios. The model have to give the number of objects of the category in the instruction.

Security. The security task evaluates the model's ability to recognize multiple types of objects in infrared images. The model have to find the object that do not exist in the image from four candidates.

Locate. The locate task evaluates the model's ability to discriminate the relationship between objects in infrared images. The model have to determine the correctness of the relative position between two specified objects.

4.3 Construct

Task	Source	Count	Percentage (%)
Locate	FLIR	4387	14.48
Aerial count	Vedai, Visdrone	3721	16.42
Pedestrian count	LLVIP	2388	10.54
Recognition	LSOTB	7463	32.94
Scene	RGBNIR	477	2.1
Security	FLIR	4219	18.62
Total		22655	100.00

Table 1: **Task Counts and Percentages.** Based on seven different thermal datasets, we constructed six tasks—localization, aerial counting, pedestrian counting, recognition, scene analysis, and security—resulting in a total of 22,655 question-answer pairs.

Template-based construction. We constructed the infrared template instruction benchmark based on real infrared images and their annotations, combined with different templates (Dai et al., 2024), as shown in Figure 2. With the real annotations, the accuracy and consistency of the benchmark can be ensured. We construct a total of 22,655 question-answer pairs. The number and task composition of the benchmark are shown in the Table 1. Please refer to Appendix B for more details.

5 Infrared-LLaVA

Given the lack of a MLLM trained on infrared data, we construct an infrared-specific MLLM, Infrared-LLaVA-7B, that utilizes infrared data for a two-stage training process involving alignment and supervised fine-tuning.

5.1 Model Architecture

The Infrared-LLaVA-7B, following the design of LLaVA 1.5 (Liu et al., 2023), comprises an infrared encoder, a trainable align layer, and a base language model. Specifically, we adopt Language-Bind Infrared Encoder (Zhu et al., 2023a) as the infrared encoder, a multi-layer perceptron(MLP) as the align layer, and the Vicuna1.5-7B (Chiang et al., 2023) as the base language model, as shown in Figure 3. Infrared-LLaVA-7B uses LanguageBind-Infrared Encoder to extract infrared features from infrared images, and then aligns them to the text embedding space through the MLP. Then

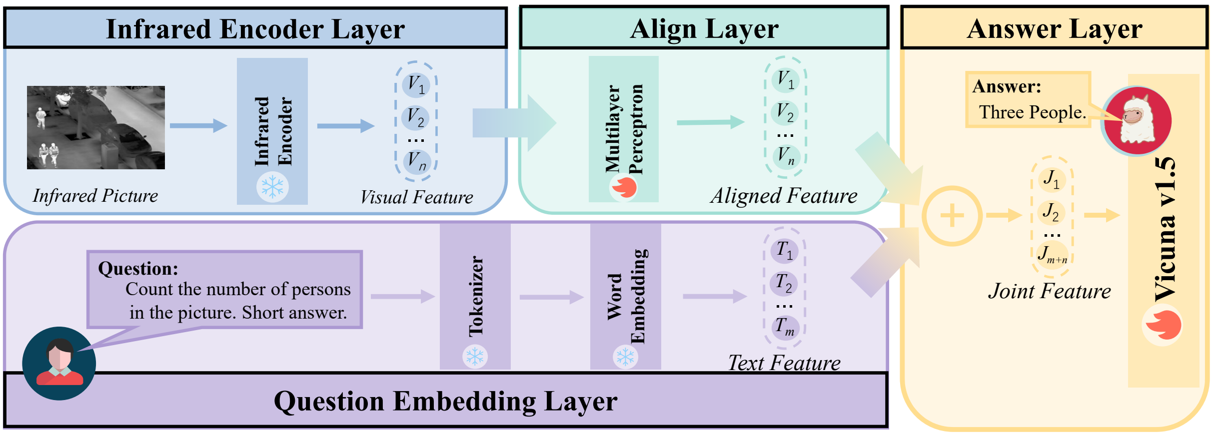


Figure 3: **Infrared-LLaVA Framework.** The Infrared-LLaVA framework shows the data flow that generates corresponding responses based on infrared image and input instruction.

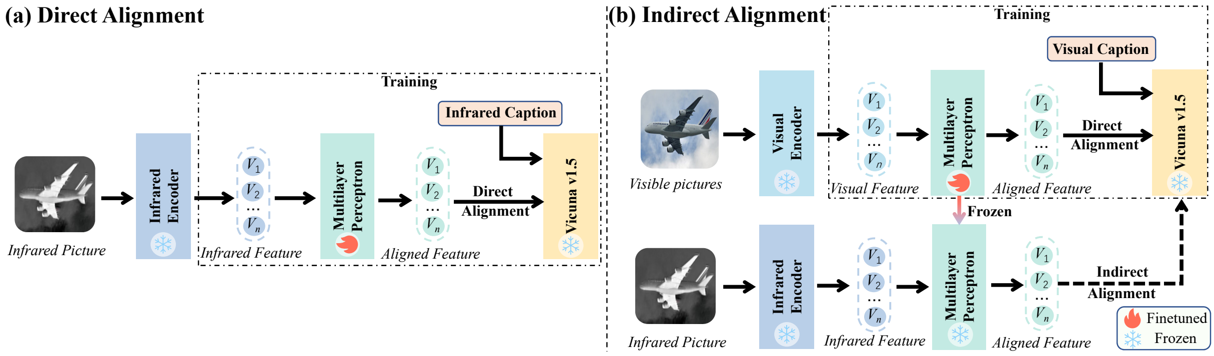


Figure 4: **Direct Alignment and Indirect Alignment.** The left image shows the process of using infrared image-text pairs to train the alignment layer and achieve direct alignment, while the right image shows the process of using visible image-text pairs to train the alignment layer and achieve indirect alignment.

the aligned features and text embedding features are concatenated to generate text using the base large language model.

5.2 Alignment

Based on the LanguageBind's shared embedding space (Zhu et al., 2023a; Zhang et al., 2023), we train only on visible image-text pairs to achieve the indirect alignment between infrared images fartures embedding space with the LLM's text embedding space. We refer to this approach as indirect alignment, distinguishing it from direct alignment method that utilizes infrared image-text pairs for alignment. To explore their impact on model ability, we implemented the two methods.

Direct Alignment. In direct alignment, we utilize the 500k infrared image-text pairs from 3.2 and used the LanguageBind's Infrared Encoder to extract infrared image features for training to align the infrared image feature space with the text embedding space directly, as shown in Figure 4(a).

Indirect Alignment. In indirect alignment, we utilized 500k visible image-text pairs from Liu et al. (2024). The LanguaugeBind Image Encoder is used to extract visible image features for training to align the visible image feature space with the text embedding space. Thanks to LanguageBind's shared embedding space, the alignment layer trained by visible image can indirectly align the infrared image feature space with the text embedding space, as shown in Figure 4(b).

In both alignment processes, the paired texts are converted into the instruction format to compute the infrared-condition text generation loss. All modules except the alignment layer are frozen.

5.3 Supervised Fine-tuning

The supervised fine-tuning (SFT) stage aims to improve the model's responsiveness to infrared-related instructions based on alignment. We use the 12k infrared instruction data obtained in 3.3 as training data. We freeze the encoder and fine-tune the alignment layer and LLM, and also minimize the infrared-condition text generation loss.

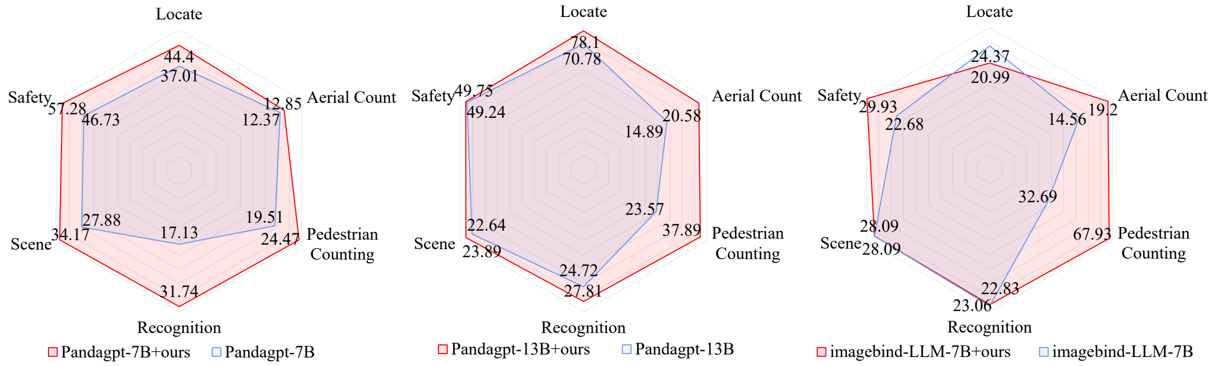


Figure 5: **Verification of the effectiveness of infrared instructions in the SFT stage.** We compared the performance of the three models on the Infrared Template Benchmark before and after SFT with our infrared instructions.The infrared instruction data is only equivalent to 7.5% of the SFT data of Pandagpt-7B and Pandgpt-13B and 6% of the SFT data of imagebind-LLM-7B instructions.

Model	Locate	Aerial Count	Pedestrian Counting	Recognition	Scene	Safety	Avg
(a) Zero Shot							
ImagebindLLM-7B	24.37	14.56	32.69	23.06	28.09	22.68	24.24
Pandagpt-7B	37.01	12.37	19.51	17.13	27.88	46.73	26.77
Pandagpt-13B	70.78	14.89	23.57	24.72	22.64	**49.24**	34.31
Infrared-LLaVA-7B	**78.10**	**20.42**	**68.90**	**64.71**	**68.55**	38.67	**56.57**
(b) Fine-tuning							
ImagebindLLM-7B-FT	81.69	18.87	50.89	30.32	31.86	77.38	48.50
Pandagpt-7B-FT	80.17	16.19	16.87	39.82	27.88	66.97	41.31
Pandagpt-13B-FT	78.17	16.19	25	43.97	24.73	88.01	46.01
Infrared-LLaVA-7B-FT	**82.59**	**35.63**	**85.50**	**95.85**	**81.76**	**94.54**	**78.98**

Table 2: **Comparison with other models.** We present the scores and average scores of Infrared-LLaVA and the other three models on each task, including zero-shot testing and fine-tuning testing. The best results for each task are highlighted in **bold**.

6 Experiments

6.1 Benchmark and Evaluation Metrics

We use the Infrared Template Benchmark constructed in Section 4 to evaluate the understanding ability of infrared images. The data from this benchmark is split into train and test subsets in 2:1. In the main experiments, we use the accuracy of each task, such as "Locate Acc," as the evaluation metric. In the ablation experiments, we use the average accuracy of each task "ITB Avg" as the evaluation metric.

6.2 Dataset Verification

We further fine-tune the existing multi-modal large model that can encode infrared images using the 12k infrared instruction in the Infrared Instruction Dataset to verify the effectiveness of the generated infrared instructions. The models we selected include imagebind-LLM-7B, Pandagpt-7B, and Pandagpt-13B. As shown in Figure 5, the performance of the models fine-tuned with the Infrared Instruction Dataset is improved compared to the original models, indicating that the generated data can effectively enhance the model's understanding ability for the infrared images and instructions.

6.3 Comparison with Other Model

Based on the validity of the verification data, we compared our proposed Infrared-LLaVA-7B with other models. It should be noted that Infrared-LLaVA-7B in Section 6 is aligned by the direct alignment method.

Zero-shot reasoning. The selected comparison models are trained using visible image-text pairs and visible image instructions. Similarly, **Infrared-LLaVA-7B**, which is trained only generated infrared data, does not use any infrared template instruction data during training. Therefore, the six template instruction tasks tested are **unseen** to these models. As shown in Table 2(a), our Infrared-

Align Method	SFT data	ITB Avg
direct	visual image instruction	47.11
indirect	visual image instruction	52.57
indirect	infrared instruction	55.23
direct	infrared instruction	**56.57**

Table 3: **Ablation on the alignment method.** In the alignment stage, the direct alignment and indirect alignment methods are used. In the SFT stage, 12k infrared instructions and 12k visual image instructions are used.

Align Method	SFT structure		ITB Avg
	Train MLP?	Train LLM?	
direct	✓	✗	44.28
direct	✗	✓	54.54
direct	✓	✓	**56.57**

Table 4: **Ablation on the SFT structure.** Results of different SFT structures on Infrared Template Benchmark based on direct alignment and infrared instructions.

Align Method	Align Structure	ITB Avg
direct	Qformer	44.28
direct	Linear	55.20
direct	MLP	**56.57**

Table 5: **Ablation on the align structure.** In addition to the MLP used in this paper, the same linear structure as MiniGPT-4 (Zhu et al., 2023b) and the same qformer structure as BLIP2 (Li et al., 2023a) are used to perform the same alignment and SFT process.

LLaVA-7B achieves the best performance on five tasks in Infrared Template Benchmark during zero-shot reasoning, especially in pedestrian counting, recognition, and scene recognition, which are 45, 40, and 46 percentage points higher than Pandagpt-13B. This shows that our generated infrared data helps to enhance the model's ability to respond to unseen thermal question answering tasks.

Fine-tuning on the benchmark. The above models are further fine-tuned using the training set of Infrared Template Benchmark. As shown in Table 2(b), the fine-tuned Infrared-LLaVA-7B achieves the best performance in all tasks. This further demonstrates that our generated data helps the model understand infrared images and respond to related task instructions.

6.4 Ablations

Indirect alignment and direct alignment. As shown in Table 3,the results on Infrared Template Benchmark after SFT with infrared instructions based on direct and indirect alignment differ by only 1.3. This proves that the modality alignment achieved on the direct alignment and indirect alignment is comparable. Meanwhile, with the same alignment method, using visible image instructions leads to decreases of 2.7 and 9.4 compared to using infrared instructions in the SFT stage. This may indicate using infrared instructions in the STF stage can reduce the model's biased understanding of infrared images compared to using visible images instructions.

The impact of SFT structure. As shown in Table 4, training the alignment layer and LLM at the same time has an improvement of 12.29 and 2.03 compared with only training the alignment layer or only training the LLM. This suggests that updating the parameters of the alignment layer and LLM in the SFT stage is more effective in enhancing the large model's ability to understand and respond to infrared images and instructions.

Type of align layers. As shown in Table 5, with the same data composition and SFT settings, using a MLP as the align layer surpasses using a qformer with 12.3 and using a linear layer with 1.3. This demonstrates that the alignment layer being too simple or too complex reduce the modality alignment effect. Using a MLP may be a better choice for modality alignment.

7 Conclusion

In this paper, we propose an debating multi-agent data generation system, which includes a Caption Agent that generates infrared images and captions, a Task Agent that generates instructions, and a Debate Agent that evaluates the instructions. Using this system, we transfer the visible image dataset COCO to get 118k infrared images and 500k infrared image-text pairs, as well as 12k infrared instruction. We validate the effectiveness of this dataset through incremental fine-tuning on other MLLMs. We further construct the Infrared-LLaVA-7B, which is fully aligned and fine-tuned based on infrared data, demonstrating the importance of infrared instruction data. We explore the impact of alignment methods, SFT structures, and alignment layer types on the model's infrared question-answering performance using Infrared-LLaVA-7B.

Limitations

Noise and hallucination. Although the instructions generated by the multi-agent system perform well during actual training, there are still some issues with noise and hallucinations. Since GPT-3.5 is a large language model trained on general-purpose corpora, the training data might lack specialized knowledge texts related to infrared images, potentially leading to some noise and inaccuracies when generating or interpreting instructions. And part of the hallucinations comes from the inherent hallucinations of GPT-3.5 itself, while another part arises from the fact that GPT-3.5 has not actually seen infrared images. Therefore, in future work, it is worthwhile to explore ways to reduce noise in the generation process by further fine-tuning the model with relevant texts from the infrared domain, as well as researching how to minimize hallucinations with a multi-modality large language model which can understand infrared image.

Incomplete test benchmarks. The benchmark we constructed based on common infrared datasets still has room for improvement. Our benchmark is limited to 6 common tasks, with poor diversity in instructions. The types of tasks and numbers of instructions need to be further expanded. Additionally, this benchmark only considers infrared tasks in general scenarios and does not account for application scenarios in fields like industry and medicine. Tasks such as temperature measurement and component damage detection need to be included to evaluate the model's understanding of infrared images in specialized scenarios.

Further validation of alignment. We only explored the comparison between indirect alignment using visual images and direct alignment using infrared images. Further validation of the indirect alignment effect based on a unified embedding feature space is needed across more modalities, including X-ray images, depth images, and etc.

Effectiveness of each agent. We only verified the effectiveness of data generated by the whole system. The further ablation study is needed to assess the necessity of each agent in the constructed debating multi-agent system.

Acknowledgments

The research in this article is supported by the National Key Research and Development Project (2021YFF0901602), the National Science Foundation of China (U22B2059, 62276083). Besides, We would like to express our gratitude to Du Xiaoman Technology for their support of our research paper.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Matthew Brown and Sabine Süsstrunk. 2011. Multi-spectral sift for scene category recognition. In *CVPR 2011*, pages 177–184. IEEE.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Chenjun Gao, Yanzhi Dong, Xiaohu Yuan, Yifei Han, and Huaping Liu. 2023. Infrared image captioning with wearable device. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8187–8193. IEEE.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190.

Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, Xudong Lu, Shuai Ren, Yafei Wen, Xiaoxin Chen, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. Imagebind-llm: Multi-modality instruction tuning. *Preprint*, arXiv:2309.03905.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. *Advances in neural information processing systems*, 27.

Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. 2021. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3496–3504.

Alex Kim, Keonwoo Kim, and Sangwon Yoon. 2024. Debate: Devil's advocate-based assessment and text evaluation. *arXiv preprint arXiv:2405.09935*.

Dong-Guw Lee, Myung-Hwan Jeon, Younggun Cho, and Ayoung Kim. 2023. Edge-guided multi-domain rgb-to-tir image translation for training vision tasks with challenging labels. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8291–8298. IEEE.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023b. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2023. Vila: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Qiao Liu, Xin Li, Zhenyu He, Chenglong Li, Jun Li, Zikun Zhou, Di Yuan, Jing Li, Kai Yang, Nana Fan, et al. 2020. Lsotb-tir: A large-scale high-diversity thermal infrared object tracking benchmark. In *Proceedings of the 28th ACM international conference on multimedia*, pages 3847–3856.

OpenAI. 2024. Chatgpt: Gpt-4. https://www.openai.com/chatgpt. Accessed: 2024-06-05.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. In *Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants!*, pages 11–23.

Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. 2022. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. 2023. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*.

Minghao Wu, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. 2024. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *arXiv preprint arXiv:2405.11804*.

Ruixin Yang, Dheeraj Rajagopa, Shirley Anugrah Hayati, Bin Hu, and Dongyeop Kang. 2024. Confidence calibration and rationalization for llms via multi-agent deliberation. *arXiv preprint arXiv:2404.09127*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. 2023a. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023b. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.

Junda Zhu, Lingyong Yan, Haibo Shi, Dawei Yin, and Lei Sha. 2024. Atm: Adversarial tuning multi-agent system makes a robust retrieval-augmented generator. *arXiv preprint arXiv:2405.18111*.

A Analysis of existing thermal data

Based on existing research, we collected 62 datasets including LLVIP and FLIR and conducted some analysis. As shown in Figure 6, the tasks of the collected datasets include target tracking, pedestrian detection, target classification, etc. The annotated object categories range from 1 to 47, and the shooting scenes include indoor, outdoor, aerial and sea. At the same time, due to the differences in acquisition equipment and imaging forms, the collected infrared images can be divided into three categories: grayscale images, pseudo-color images and grayscale unbalanced images. According to the data type, it can be divided into four categories: sequence, single image, video, and synthetic image. It can be seen that infrared data has the characteristics of a single modality and fewer tasks.

B Dataset Examples

We show examples of the Infrared Template Benchmark and Infrared Instruction Dataset, as shown in Figure 7 and Figure 8. It is evident that the Infrared Template Benchmark covers multiple common infrared scenarios and various types of objects, but the diversity of the instructions is relatively low. On the other hand, the instruction data quality in the Infrared Instruction Dataset is relatively high, though it still contains some hallucinations, such as the Q&A about protective gear in the "Multi Conversations Example".

C Prompts

We show prompts for the GPT-3.5 in the multi-agent system: the prompt used to filter the caption in Figure 9; the prompt used to generate three types' instructions in Figure 10, Figure 11 and Figure 12; the prompt used to judge instructions in Figure 13.

D Training Details

In the alignment stage, we train our models for 1 epoch with a learning rate of 1e-3 and a batch size of 128 with 500k generated infrared image-text pairs. In the SFT stage, we train our models for 3 epochs with a learning rate of 2e-5 and a batch size of 64. And in both stages, we use the Adam optimizer with no weight decay and a cosine learning rate with a warmup radio of 3%. To save GPU memory, the Full Shared Data Parallel(FSDP) and gradient checkpointing is used.

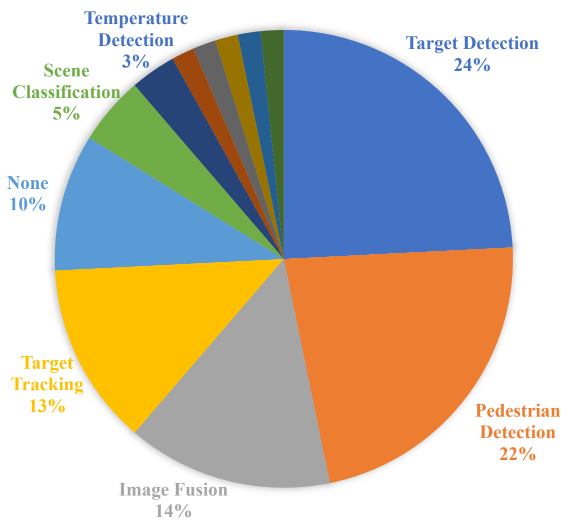
We train all models with 4xA800s. Alignment training on 500k generated infrared image-text pairs within 3.5 hours. STF on 12k generated infrared instructions with 1.5 hours.

E Potential Risks

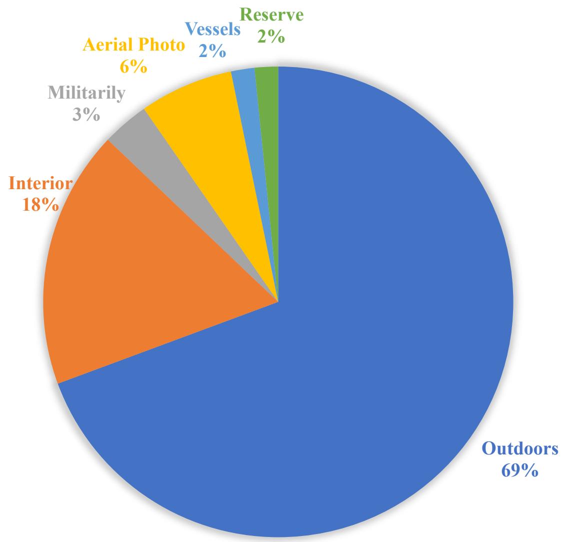
Except the hallucination and limited evaluation discussed in limitations, we discuss other potential risks in the following.

Biases. Bias can be transferred from the vision encoder (LanguageBind Infrared Encoder) and the language decoder (Vicuna/GPT-3.5). This may lead to biased outcomes or unfair representations of diverse content in the generated infrared data and the outcome of Infrared-LLaVA-7B.

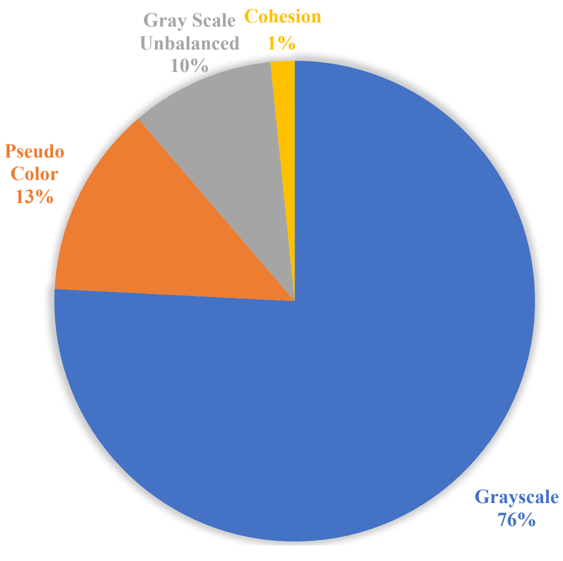
Energy consumption. Though energy consumption is not a primary concern for our work due to the smaller training data and model size than other researches, it may become a concern when scaling up the training dataset or increasing the model size.



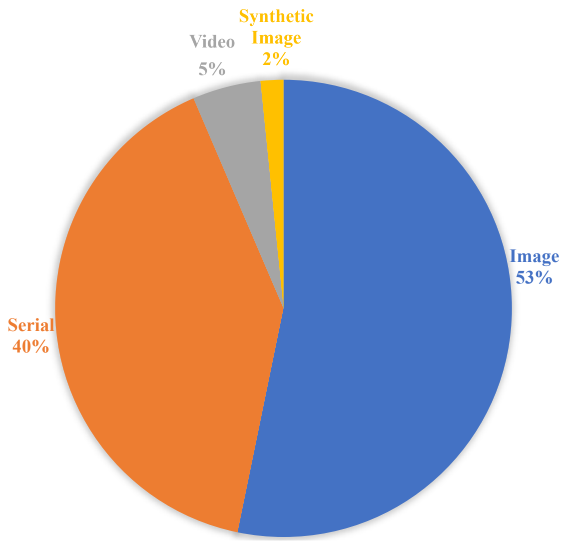
(a) Task Distribution



(b) Scene Distribution



(c) Image Type



(d) Data Type

Figure 6: **Analysis of existing infrared image data.** The four pie charts show the scene distribution, task distribution, imaging type, and data type of existing infrared datasets.

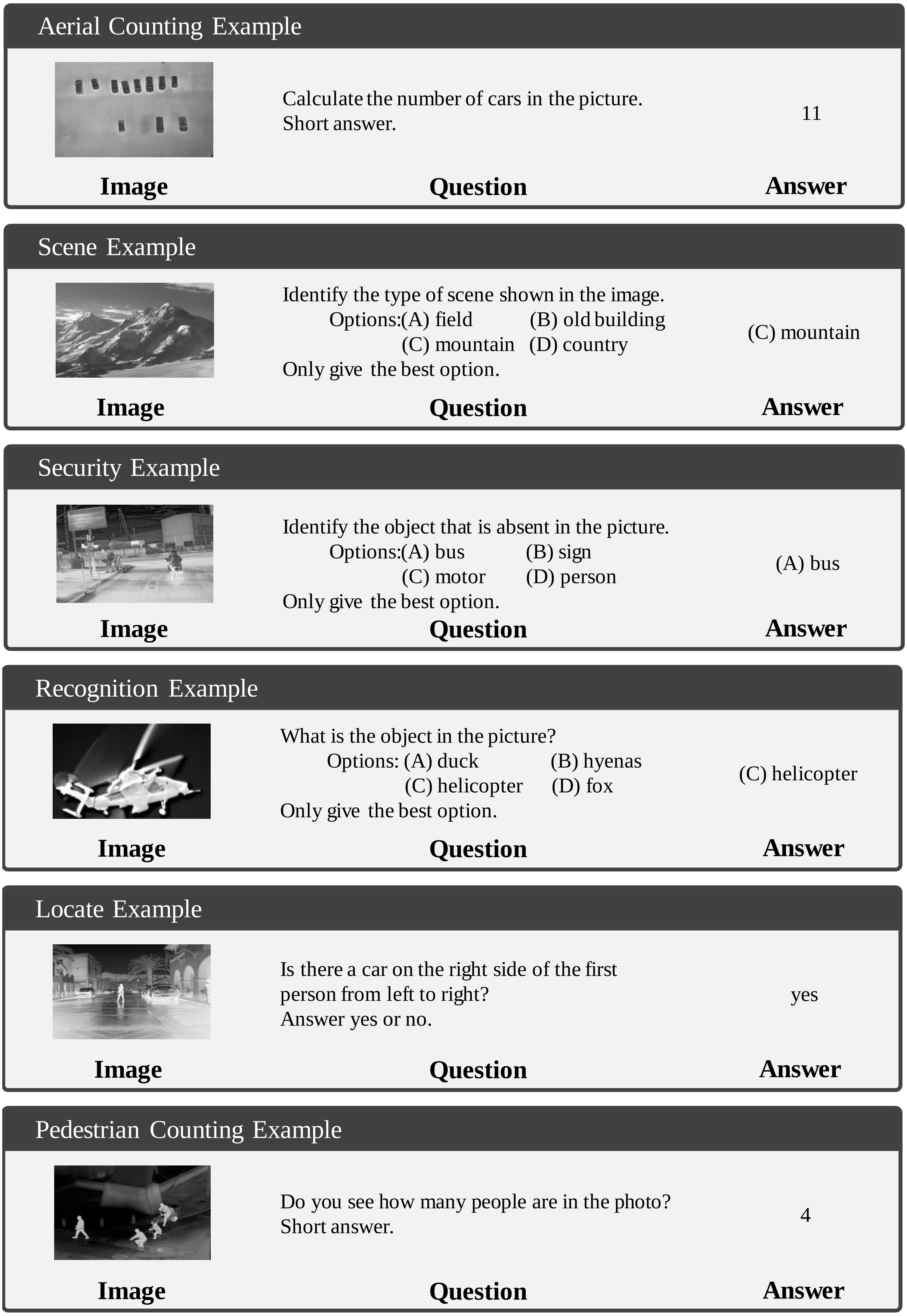


Figure 7: **Template instruction task category examples.** Image input, instruction input, and expected answers for six tasks.

Multi Conversations Example

Human :

What is the main object in the image?

GPT :

The main object in the image is a motorcycle, with a person riding it.

Image

Human :

What is the person on the motorcycle doing?

GPT :

The person on the motorcycle is performing a wheelie.

Human :

Is the person wearing any protective gear?

GPT :

Based on what I can see in the image, it is not possible to definitively determine whether the person is wearing any protective gear.

Detailed Description Example

Human :

Write an exhaustive depiction of the given image.

GPT :

Image

The image shows a group of people gathered around two motorcyclists in uniform. The motorcyclists are driving between the group of people, with one motorcycle on the left side of the scene and the other on the right. Overall, it appears to be a police escort or motorcade, with the motorcyclists leading the way and the group of people following behind.

Complex Reasoning Example

Human :

Image

What is the significance of the clock tower?

GPT :

The clock tower is significant because it serves as a landmark in the town square. It is likely that the clock tower is a historical monument that has been standing for a long time, and it is a symbol of the town's heritage.

Figure 8: **Infrared instruction dataset examples.** Synthetic infrared instruction examples for three tasks.

Caption Agent Prompt

You are an AI visual assistant capable of analyzing a single infrared image caption. You will receive several sentences describing the same infrared image, along with specific infrared feature filtering rules. Your task is to filter out words from the given caption that do not belong to the infrared feature and ensure that the final output is fluent and natural. Do not directly mention the filtering rules. Instead, use natural language to describe the filtered content. The final description should be complete, coherent, and accurately reflect the scene in the infrared image.

User:

Caption:
{caption}
Rules:
{rules}

Assitant:

Example

User:

Caption:
A big white steel plane flies in the blue sky.
Rules:
Delete all color related words.
Delete all transparency related words.
Delete all texture related words.
Delete all transparency related words.
Delete all brightness related words.

Assistant:

Filter Caption:
A big steel plane flies in the sky.

Figure 9: The Caption Agent prompt demonstration.

Task Agent Prompt (Multi Conversation)

You are an AI visual assistant, and you are seeing a single infrared image.In addition, specific object locations within the infrared image are given, along with detailed coordinates. These coordinates are in the form of bounding boxes, represented as (x1, y1, x2, y2) with floating numbers ranging from 0 to 1. These values correspond to the top left x, top left y, bottom right x, and bottom right y. What you see are provided with several sentences and these coordinates, describing the same infrared image you are looking at. Answer all questions as you are seeing the infrared image.
Design a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the infrared image and answering the question.
Ask diverse questions and give corresponding answers.
Include questions asking about the visual content of the infrared image, including the object types, counting the objects, object actions, object locations, relative positions between objects, etc. Only include questions that have definite answers:
(1) one can see the content in the infrared image that the question asks about and can answer confidently;
(2) one can determine confidently from the infrared image that it is not in the infrared image.
Do not ask any question that cannot be answered confidently.
Also include complex questions that are relevant to the content in the infrared image, for example, asking about background knowledge of the objects in the infrared image, asking to discuss about events happening in the infrared image, etc. Again, do not ask about uncertain details.
Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary.

User:

Captions:
{captions}
Boxes:
{bboxes}
Specific Requirements:
{requirements}

Assitant:

Example

User:

Captions:
There is a movie theater that displays the show times.
A fire hydrant is deep in the snow.
The fire hydrant is in the snow near a sidewalk.
A hotel for dogs in the snow in winter.
Boxes:
fire hydrant: [0.326, 0.612, 0.426, 0.72]
Specific Requirement:
Material information is not visible in the IR image.

Assistant:

Question:What is the main object in the image?
Answer:The main object is a fire hydrant in the image.
===
Question:Is the fire hydrant fully visible or partially visible?
Answer:The fire hydrant is partially visible, with a significant portion of it covered by something.
===
Question:What is the condition of the sidewalk near the fire hydrant?
Answer:The sidewalk near the fire hydrant has recently been plowed, so it looks clearer here compared to the surrounding area.
===

Figure 10: The Task Agent generates the multi-conversations task prompt demonstration.

Task Agent Prompt (Complex Reasoning)

You are an AI visual assistant that can analyze a single infrared image. You receive several sentences, each describing the same infrared image you are observing. In addition, specific object locations within the infrared image are given, along with detailed coordinates. These coordinates are in the form of bounding boxes, represented as (x1, y1, x2, y2) with floating numbers ranging from 0 to 1. These values correspond to the top left x, top left y, bottom right x, and bottom right y.
The task is to use the provided caption and bounding box information, create a plausible question about the infrared image, and provide the answer in detail.
Create complex questions beyond describing the scene.
To answer such questions, one should require first understanding the visual content, then based on the background knowledge or reasoning, either explain why the things are happening that way, or provide guides and help to user's request. Make the question challenging by not including the visual content details in the question so that the user needs to reason about that first.
Instead of directly mentioning the bounding box coordinates, utilize this data to explain the scene using natural language. Include details like object counts, position of the objects, relative position between the objects.
When using the information from the caption and coordinates, directly explain the scene, and do not mention that the information source is the caption or the bounding box. Always answer as if you are directly looking at the image.

User:

Captions:
{captions}
Boxes:
{bboxes}
Specific Requirements:
{requirements}

Assitant:

Example

User:

Captions:
A man in a shirt wearing very many ties.
a man with ties on poses for a picture
A man wearing multiple ties on his neck.
A man smiles while wearing several ties.
Boxes:
tie: [0.574, 0.298, 0.752, 0.704]
tie: [0.464, 0.339, 0.639, 0.789]
person: [0.019, 0.065, 0.962, 0.988]
tie: [0.44, 0.318, 0.63, 0.397]
tie: [0.566, 0.33, 0.655, 0.543]
tie: [0.25, 0.409, 0.359, 0.556]
Specific Requirement:
Material information is not visible in the IR image.

Assistant:

Question:
What is unusual about this photo?
Answer:
In the infrared photo, the man is wearing so many ten ties around his neck. This is considered unusual because it is not common for someone to wear multiple ties at once. In a typical setting, a person would wear only one tie.

Figure 11: The Task Agent generates the complex reasoning task prompt demonstration.

Task Agent Prompt (Detailed Description)

You are an AI visual assistant that can analyze a single infrared image. You receive several sentences, each describing the same infrared image you are observing. In addition, specific object locations within the infrared image are given, along with detailed coordinates. These coordinates are in the form of bounding boxes, represented as (x1, y1, x2, y2) with floating numbers ranging from 0 to 1. These values correspond to the top left x, top left y, bottom right x, and bottom right y.
Using the provided caption and bounding box information, describe the scene in a detailed manner. Instead of directly mentioning the bounding box coordinates, utilize this data to explain the scene using natural language. Include details like object counts, position of the objects, relative position between the objects.
When using the information from the caption and coordinates, directly explain the scene, and do not mention that the information source is the caption or the bounding box. Always answer as if you are directly looking at the infrared image.

User:

Captions:
{captions}
Boxes:
{bboxes}
Specific Requirements:
{requirements}

Assitant:

Example

User:

Captions:
A harbor filled with lots of boats next to a building.
A bicycle parked in front of several boats at a dock.
A bicycle in front of a line of docked white yachts
A bike sits before boats which sit before a long building.
Boxes:
bicycle: [0.287, 0.641, 0.507, 0.874]
bicycle: [0.566, 0.667, 0.63, 0.731]
boat: [0.541, 0.611, 0.668, 0.731]
person: [0.778, 0.527, 0.797, 0.57]
cup: [0.708, 0.733, 0.724, 0.758]
boat: [0.957, 0.526, 1.0, 0.752]
Specific Requirement:
Material information is not visible in the IR image.

Assistant:

It is a harbor filled with numerous boats of various sizes docked next to a long building. Among the boats, there are a few yachts lined up, standing out from the rest. There is a bicycle prominently parked in front of the line of docked boats, serving as a convenient means of land transportation for those living on the boats. Another bicycle can be seen further back in the scene, near the middle of the harbor.

A person is visible near the right side of the harbor, possibly enjoying the view or attending to their boat. Additionally, there is a cup placed on a surface near the middle of the scene.

Figure 12: The Task Agent generates the detailed description task prompt demonstration.

Debate Agent Prompt

You are an AI visual assistant responsible for evaluating the instruction data related to infrared images. You will receive an instruction data that is related to an infrared image. Your task is to determine if there are any issues with this instruction data. If you find any problems, clearly point them out. If there are no issues, simply respond with "ACCEPT". Always ensure your assessment is accurate and thorough, addressing any potential ambiguities or inconsistencies in the instruction data. Do not mention that your source of information is the instruction data itself. Always respond as if you are directly analyzing the task requirements.

User:

Instruction Data:

{instruction}

Assitant:

Examples of rejection

User:

Instruction Data:

Question:

What material is the plane made of ?

Answer:

The plane is made of steel.

Assistant:

Material information is not visible in the infrared image.

Examples of acceptance

User:

Instruction Data:

Question:

Where is this plane flying?

Answer:

It's flying in the sky.

Assistant:

ACCEPT

Figure 13: The Debate Agent prompt demonstration.