基于大模型的多模态视觉理解与推理

一、赛题背景

本赛题面向新一代人工智能理解与推理能力的发展需求，围绕语言与多模态视觉联合理解问题展开。传统目标检测只需识别图像中的物体类别并给出其位置，而真实智能体则需具备更高层次的认知能力：理解人类的自然语言描述，并在复杂场景中精准定位所指目标，该问题被定义为视觉定位（Visual Grounding）。大语言模型、多模态大模型与推理大模型的发展，使人工智能逐步具备从语义理解、跨模态关联到复杂场景决策的综合能力。这些模型正在推动人工智能从“感知世界”迈向“理解意图并做出决策”的阶段。然而，在真实环境中，仅依赖单一视觉信息往往难以稳定定位目标，尤其在弱光、遮挡或外观相似物体存在时，对语义与空间一致性的综合理解能力提出了更高要求。本赛题在视觉模态上引入可见光、深度与热红外三种视觉模态，参赛者需要针对复杂光照、遮挡等环境数据完成跨模态语义对齐与推理。

本赛题提供空间对齐的可见光、深度与热红外图像及文本指令描述，参赛者需基于大模型技术并设计算法，实现由语言到空间位置的映射，输出目标区域位置坐标。出题思路旨在推动从感知型 AI 向理解型 AI 转变，鼓励参赛者探索大语言模型与视觉模型的协同机制，例如跨模态注意力、语义对齐、推理增强与外部知识辅助等方法，从而提升模型在真实环境中的理解与决策能力。本赛题重点考察参赛者在多模态表示学习、语言视觉对齐、跨模态推理、复杂环境鲁棒感知以及系统实现能力等方面的综合水平，引导学生理解从检测物体到理解意图并找到物体的智能感知范式转变，培养面向新一代通用人工智能的重要基础能力。

二、赛题应用场景

基于大模型的多模态视觉理解与推理，对于自动驾驶、智能安防、应急救援、工业巡检等诸多实际应用场景都具有重要应用价值。其中文本模态作为语义引导核心，与红外、可见光、深度三种视觉模态协同，是实现“语义描述-视觉目标”精准匹配的关键。例如，在智能安防应急处置场景中，当监控后台接到报警指令，需根据语义描述（如“园区西北角围墙处，蹲在草丛中、身高约 1.7 米的可疑人员”）这一文本模态输入，在复杂监控画面（红外、可见光、深度视觉模态）中精准定位目标时，系统需融合可见光 RGB 摄像头（捕捉目标衣着纹理、草丛及围墙外观细节）、红外热成像传感器（弥补低光缺陷、捕捉目标热源特征，区分活体与静物）、深度传感器（获取目标三维尺寸、与围墙的空间距离，匹配文本描述中的身高信息）三类视觉异构数据，结合文本模态提供的语义线索，由大模型完成文本语义与多模态视觉特征的精准对齐、关联推理，实现目标的快速定位。当目标因可见光画面无法清晰分辨，或夜间光线昏暗导致可见光成像模糊时，红外热成像可清晰突显目标热源轮廓，深度数据可通过三

维几何信息验证目标身高、空间位置是否与文本描述匹配，文本模态则提供明确的语义约束，排除无关背景干扰；大模型进一步挖掘文本语义与多模态视觉特征的深层关联，精准锁定符合文本描述的目标区域。该场景能够直观体现红外、可见光、深度视觉模态与文本模态协同应对遮挡、弱光、背景复杂等问题的核心作用。

三、赛题任务

本赛题旨在利用大模型多模态视觉理解与推理技术，实现复杂场景下视觉定位的精准高效完成，即对视觉三模态与查询文本模态输入进行协同处理分析，依托文本语义引导，利用多模态视觉数据的互补优势，实现给定语义描述对应的目标精准定位。

（一）任务描述

在多模态视觉定位任务中，文本模态作为核心语义引导，给定目标的文本语义描述（描述目标表观特征、空间位置或行为状态），参赛者需要对可见光 RGB 图像、红外图像、深度图像三种视觉模态数据，以及文本模态数据进行协同分析和处理，从而精准定位出目标在图像中的具体位置，实现文本语义描述与视觉目标的精准匹配。任务给定目标的文本语义描述，要求参赛者设计基于大模型的多模态算法，对三种视觉模态数据和文本模态数据进行协同处理、特征对齐与语义推理，实现在图像中的精准定位，如图 1 所示。

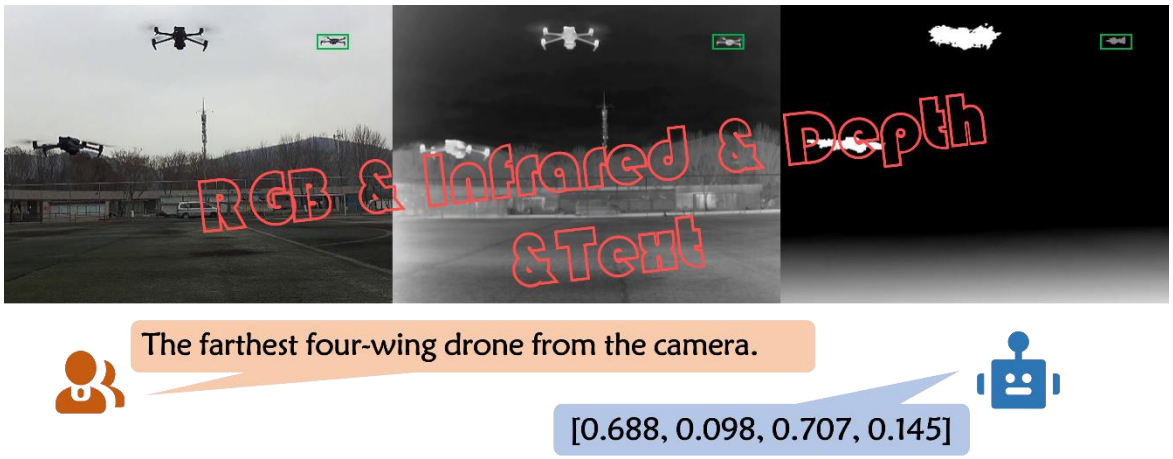

图 1 基于大模型的多模态视觉理解与推理的任务描述

（二）任务输入输出说明

1. 数据集样本说明：一个数据集样本包含四个模态——可见光模态、热红外模态、深度模态、文本模态。其中，可见光模态像素值对应 RGB 三通道红、绿、蓝色彩强度；热红外模态像素值对应物体表面辐射温度大小；深度模态像素值表示相机到对应目标点的空间距离；文本模态为一个描述特定目标的英文句子。

2. 数据集标签说明：每个样本的标签为文本对应目标在可见光图像中的边界框，格式为[x1, y1, x2, y2]，且所有坐标均经过归一化处理（即输出为图像尺寸的相对值，取值范围为[0, 1]）；其中 x1、y1 为目标边界框左上角相对坐标，x2、y2 为目标边界

框右下角相对坐标，坐标相对值由像素绝对值除以图像对应边长（宽度对应 x 轴、高度对应 y 轴）得到；标签与文本模态描述一一对应，用于验证及定位结果的精准评估。

算法输入：可见光图像（RGB）、深度图（Depth）、热红外图像（Infrared），以及目标语义描述查询文本模态数据（Query，描述目标的英文句子）。

算法输出：目标在输入的可见光图像中的位置，即可见光图像中目标的边界框（Bounding Box, BBox），格式为[x1, y1, x2, y2]，且所有输出坐标均需经过归一化处理（即输出为相对值，取值范围为[0, 1]）。

四、数据集及数据说明

（一）数据来源

本赛题提供的数据集为自行采集的数据，查询文本描述是对目标的外观、形状、位置等的描述，由人工进行文本标注，文本描述确保仅对应图像中的唯一目标。赛题所提供数据集在采集过程中涉及到的受试者均获得了同意。

（二）数据内容

赛题数据共包含初赛、复赛、半决赛评测集，由可见光-红外-深度三模态图像组构成。一组图像组对应一个或多个查询文本（Query），每条文本指向可见光图像中唯一的目标。

（三）数据格式

视觉信息包含可见光、红外、深度三模态数据。其中，可见光为三通道 8 位无符号整型数据，像素取值范围为[0, 255]。红外图像存储为三通道 8 位无符号整型数据，各通道视觉特征高度一致，无实际彩色语义，本质为 3 个单通道热辐射灰度图像堆叠，像素取值范围为[0, 255]，像素数值越高表示目标温度越高，画面视觉效果越亮。深度图像为单通道 16 位无符号整型数据，理论取值范围为[0, 65535]，单位为毫米，深度相机的可感知距离范围约为 30 厘米到 20 米（即取值范围[0, 19999]）。深度图像素值为 0 或者过小代表无效深度区域，数值越大表示目标距离相机越远，数值越小表示目标距离相机越近。所有图像均未经过归一化处理。

各模态图像的关联信息以 JSON 格式结构化存储，以唯一的查询文本 ID（如“000000_001”“000001_001”）作为索引，对应每组图像的三类模态图像存储路径（相对路径）、具体的查询文本以及该查询目标在可见光图像中的归一化边界框（BBox）坐标（四元组，范围 0~1）；

示例数据可以通过以下链接获取，正式数据将在报名后开放下载。

https://pan.baidu.com/s/1wIrUz1Zk8ptALpIZkGr_9Q?pwd=2w7f 提取码: **2w7f**

（四）数据特色

本数据集全部采集于真实场景，涵盖街道、校园、动物园、室内、公园等多元环境，能够有效检验模型的跨场景泛化性能。所有视觉模态数据均已完成人工对齐。在

文本查询设计上，数据集围绕多维度任务展开，全面考察多模态互补融合、空间关系认知、动作识别、目标计数与指代推理等核心能力。

五、算法设计要求

（一）模型类型

鼓励参赛者基于大模型技术体系，采用并深度优化适配多模态场景的深度学习算法。优先支持以大模型为核心的技术路线，包括但不限于：基于开源基座大模型（如 LLaMA-Grounding、Qwen-VL、InternVL、SAM 等）的轻量化微调、基于大模型的多模态融合算法开发、大模型推理加速与显存优化算法设计等，充分发挥大模型在语义理解、跨模态关联、复杂推理等方面的核心优势。

（二）创新性

鼓励参赛者围绕大模型的核心特性提出创新性改进方案，重点提升大模型对多模态数据的协同推理、精准理解与高效决策能力。例如，针对通用大模型在特定多模态场景下的短板（如视觉定位精度不足、跨模态语义偏差等），设计专属的视觉编码器、模态交互层或后处理模块，强化大模型对多模态数据的理解能力与定位准确性。

六、性能指标要求

本赛题采用准确率（ACC@0.5）作为评估指标，判断定位正确的依据为预测框与真实框交并比（Intersection over Union，IoU）≥ 0.5。 IoU 的具体计算方式如下：

$$\mathrm{IoU}(A,B)=\frac{|A\cap B|}{|A\cup B|},$$

其中，A 代表预测区域（预测的 BBox），B 代表真实区域（标注的真实 BBox）。$A\cap B$ 表示二者的交集区域，$A\cup B$ 表示二者的并集区域，而符号 $|\cdot|$ 表示计算该区域的面积。最终，IoU 通过交集面积除以并集面积 得出一个介于 0 到 1 之间的值，其值越大，表示预测与真实结果的重合度越高。

最后，准确率（ACC@0.5）的计算公式为：

$$\mathrm{ACC@0.5}=\frac{IoU\geq 0.5\ 的有效查询样本数量}{测试集全部查询样本总数量}。$$

七、功能要求

（一）准确性

算法模型在实现面向多源异构数据（可见光、深度、红外、文本等）的视觉定位任务时，需具备较高的准确性，确保在复杂场景（如光照变化、目标遮挡、背景干扰、外观相似目标共存等）下，仍能根据文本描述精准定位对应目标，有效减少定位偏差、

漂移等问题。

（二）鲁棒性

面对不同质量等级的可见光、深度、红外图像数据及文本查询数据，算法需能够稳定运行并输出可靠的视觉定位结果。当其中一种或多种模态数据质量较差时（如可见光图像模糊、深度图像噪声过多、红外图像对比度不足等），算法不应出现大幅性能波动，需依托多模态协同优势，持续保持对目标定位的准确性与稳定性，确保单一模态质量缺陷不影响整体定位效果。

八、开发环境

（一）软件环境

1. 编程语言

本赛题推荐使用 Python 3.8 及以上版本（建议 3.8-3.11 区间）。该版本区间的 Python 具备更稳定的语法特性、更优的性能表现，且完全兼容大模型开发所需的核心库：

- 数据处理与可视化：NumPy（1.21+）、Pandas（1.4+）、Matplotlib（3.5+）、Seaborn 等；
- 图像 / 多模态数据处理：OpenCV-Python（4.5+）、Pillow（9.0+）等；
- 大模型核心依赖：Hugging Face Transformers（4.20+）、Accelerate、Tokenizers 等。

2. 深度学习框架

- PyTorch：推荐 2.0 及以上版本（如 2.0-2.2），该版本原生支持 torch.compile 编译优化、BetterTransformer 大模型推理加速，且兼容绝大部分开源大模型；
- TensorFlow：推荐 2.10+版本（建议 2.10-2.15），适配 Keras 3.0 多后端特性，支持 TPUEstimator 适配云端 TPU 资源，满足基于 TensorFlow 生态的大模型开发需求；
- 此外，可搭配 DeepSpeed、Megatron-LM 等大模型训练框架，实现显存优化、分布式并行训练，降低大模型训练的硬件门槛。

3. 计算资源

本赛题不对参赛队伍使用的计算资源做限制，亦不提供算力支持。参赛者可使用本地工作站或云端计算平台进行开发和训练。参赛者可根据自身模型规模（如小参数量模型轻量化训练）灵活选择资源，但需保障模型训练/推理的效率和稳定性。

（二）硬件环境

本赛题对参赛队伍使用的 CPU 型号、内存大小、GPU 型号等硬件资源不做强制性要求，参赛队伍可根据自身需求自由配置。参赛者也可以使用云计算平台，例如，阿里云天池、腾讯云 TI-ONE、百度 AI Studio、火山引擎方舟平台、智谱 AI 开放平

台等。

九、成绩评价

（一）输入数据格式要求

参赛者需要正确解析赛题 JSON 文件，正确读取对应的多模态图像以及文本查询。

（二）输出数据格式要求

参赛者需要预测所有文本查询所描述目标在可见光图像中的 bbox 字段，以[x1,y1,x2,y2]格式存放，并且经过归一化。不可修改除 bbox 字段外的其他字段，否则将导致评测程序匹配失败。针对评测过程中出现的反向坐标（x1≥x2、y1≥y2）、坐标越界、数值含 NaN、无效空框等异常边界框，评测机制将直接判定为无效预测。存储预测结果 JSON 文件格式如下：

```
{
  Query ID: {
    "visible": "Images/visible/XXXXX",
    "infrared": " Images/infrared/XXXXX",
    "depth": " Images/depth/XXXXX ",
    "query": " XXXXX ",
    "bbox": [
      x1,
      y1,
      x2,
      y2
    ]
  },
                                    ……
                                    ……
                                    ……
}
```

最终提交至赛事平台的文件需将预测结果的 JSON 文件压缩成 zip 压缩包。

（三）成绩计算依据

成绩将依据性能评估指标 ACC@0.5 进行打分。

注：若参赛者提交结果低于赛事方公布的基线成绩，赛事方有权将其认定为无效成绩。

十、解题思路

（一）数据预处理

对测试集中的 RGB、红外、深度图像分别做格式适配与标准化处理，完成 RGB 图像归一化、红外图像噪声抑制、深度图像离群值剔除，保持各模态图像的空间对齐特性；对英文文本查询做 token 化、标准化处理，适配大模型输入要求。精准解析 JSON 文件，按查询文本 ID 关联多模态图像与文本查询，确保数据读取的准确性与完整性。

（二）特征提取

采用差异化预训练模型开展多模态特征表征学习，基于 CLIP 等经典视觉预训练架构，分别完成可见光图像、红外图像与深度图像的独立特征编码，有效挖掘场景外观纹理、红外辐射特征与深度结构信息。同时融合 LLaMA-Grounding、Qwen-VL、InternVL 等开源多模态大模型，增强文本深层语义理解能力，精准解析目标描述中的实体属性、空间位置关系与关键约束条件；在此基础上搭建跨模态特征映射模块，实现视觉表征与文本语义特征的维度对齐与特征空间统一。

（三）模型构建与适配

基于 PyTorch/TensorFlow 框架，搭建以开源大模型为核心的多模态模型，设计跨模态注意力、加权融合等交互模块，实现文本语义与多模态视觉特征的融合推理；利用大模型轻量化技术对预训练模型做适配性调优，优化模型对视觉定位任务的适配能力，选用 IoU 相关损失函数优化边界框预测逻辑。

（四）推理预测与结果优化

将预处理后的多模态数据输入模型，推理输出可见光图像中目标的边界框坐标，对输出结果做归一化处理，确保坐标值在[0, 1]范围内且符合[x1, y1, x2, y2]格式。可采用多模型集成推理（如不同预训练模型预测结果加权平均）提升定位准确性，对预测框做格式校验与修正，严格按要求补充至预测结果 JSON 文件的 bbox 字段，不修改其他原始信息。

十一、 赛题约束条件

（一）算法约束

❑ 禁止调用商业闭源大模型的在线推理 API（如 GPT-4o 等商业闭源模型的远程调用接口），仅允许使用开源大模型或自研大模型进行自主开发、微调与部署。

❑ 禁止采用手工标注测试集目标位置的方式生成结果，所有预测结果必须由算法模型自主推理输出，严禁任何形式的人工干预测试结果生成过程。

（二） 数据使用约束

❑ 本赛题对训练数据使用不做约束，参赛者可自由使用各类外部公开数据集进行模型训练与优化，无外部数据集使用范围限制。若使用外部数据集，需在技术报告中明确注明数据来源、使用方式及预处理流程。

❑ 参赛者仅可使用赛事官方统一提供的初赛、复赛、半决赛测试集数据进行推

理预测，严禁私自对测试集进行标注、篡改，严禁将测试集数据用于模型训练环节。

❑ 赛事提供的所有测试数据集均为赛事专用，参赛者需严格遵守数据保密要求，禁止将数据集以任何形式（包括但不限于分享、上传至公开平台、转售、开源）泄露、传播或用于非本次赛事相关的商业、科研等用途，违反者立即取消参赛资格，并追究相关责任。

十二、参考资源

（一）文献资料

[1] Zhu X F, Xu T, Pan Y, et al. Collaborating Vision, Depth, and Thermal Signals for Multi-Modal Tracking: Dataset and Algorithm[C]//The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.

[2] Chollet, Francois, and François Chollet. *Deep learning with Python*. simon and schuster, 2021.

[3] Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." European conference on computer vision. Cham: Springer Nature Switzerland, 2024.

[4] Bai, Shuai, et al. "Qwen3-vl technical report." arXiv preprint arXiv:2511.21631 (2025).

（二）在线资源

1. 图像数据处理：OpenCV 官方文档（https://docs.opencv.org/）、Pillow 官方文档（https://pillow.readthedocs.io/）

2. 大模型基座：Hugging Face Hub（https://huggingface.co/）、ModelScope 魔搭社区（https://www.modelscope.cn/）

3. Bilibili 上的“动手学深度学习 Pytorch 版”课程，讲解了如何使用 Python 语言 Pytorch 深度学习框架实现深度神经网络，课程内容的应用实践性强。

4. Coursera 「Convolutional Neural Networks for Visual Recognition」：斯坦福大学经典课程，深入讲解计算机视觉特征提取、目标检测核心技术，为多模态视觉特征编码提供理论与实践支撑。

十三、提交要求

初赛、复赛：参赛队伍需提交测试集预测结果文件。

半决赛：参赛队伍需提交算法代码、预测结果文件、模型文件与环境、技术报告。具体提交内容如下：

（一）算法代码

参赛队伍应提交完整的算法代码，包括数据预处理、模型训练、预测推理等各个环节的代码。代码需使用符合规范的 Python 语言编写，具备清晰的注释和文档说明，

以便评审人员理解和运行。

（二）预测结果文件、模型文件与环境

参赛队伍应提交测试的结果文件、训练好的模型文件（由于模型可能较大，请提供模型下载链接），并提供模型的加载和使用说明，包括所需的运行环境、依赖库等信息。模型文件应能够在指定的测试环境中正常运行并输出预测结果。

（三）技术报告

参赛队伍应提交详细的技术报告，内容包括算法设计思路、模型架构图、实验设置（如训练参数、数据增强方法等）、性能分析（对主要指标和次要指标的详细分析）以及算法的创新点和不足之处的分析。技术报告格式采用 PDF，页数不限。内容格式可参考“附件：算法赛题技术方案大纲”。

总决赛：提交内容及具体要求以组委会后续正式通知为准。

十四、其他说明

公平性：严禁任何形式的作弊行为，包括但不限于手工标注、抄袭他人代码等。一经发现，立即取消参赛资格，并追究相关责任。

知识产权：参赛者提交的作品必须为原创，未在其他比赛中获奖或公开发表。比赛主办方有权对参赛作品进行展示、宣传等相关活动，但知识产权仍归参赛者所有。

十五、联系方式

赛题交流 QQ 群：1083696760

邮　箱：xuefeng.zhu@jiangnan.edu.cn

报名官网：www.aicomp.cn