ChatENV：一种用于传感器引导的环境监测和场景模拟的交互式视觉语言模型

Hosam Elgendy, Ahmed Sharshar, Ahmed Aboeitta, and Mohsen Guizani
Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

***Abstract*—理解航空图像中的环境变化对于气候适应性、城市规划和生态系统监测至关重要。然而，目前的视觉语言模型（VLMs）忽视了环境传感器中的因果信号，依赖于易受风格偏见影响的单一来源字幕，并缺乏互动场景推理。我们提出了ChatENV，这是第一个能够共同推理卫星图像对和真实世界传感器数据的互动 VLM。我们的框架：(i) 创建了一个包含 177k 图像的数据集，形成 152k 时间对，覆盖 197 个国家的 62 个土地使用类别，并且拥有丰富的传感器元数据（例如，温度、PM $_{10}$ 、CO）；(ii) 使用 GPT-4o 和 Gemini 2.0 对数据进行注释，以增加风格和语义多样性；(iii) 使用高效的低秩适配器（LoRA）对 Qwen-2.5-VL 进行微调以用于聊天。ChatENV 在时间推理和“假设”推理（例如，BERT-F1 0.902）方面表现出色，并且可以与最先进的时间模型媲美或超越，同时支持互动场景分析。这使得 ChatENV 成为一个强大的工具，用于扎根、传感器感知的环境监测。**[1]

***Index Terms*—Environmental Monitoring, Vision-Language Models, Scenario Prediction, Remote Sensing.**

I. 介绍

监测和理解环境动态是解决气候变化、城市发展和栖息地退化等紧迫问题的首要任务。遥感，特别是通过航空和卫星成像的使用，提供了地球表面的可扩展监测，使得对土地覆盖、植被、城市区域及其他环境指标的变化进行实时检测成为可能 [1], [2] 。近年来，视觉语言模型（VLMs）因其融合视觉和文本数据的能力而备受关注，使得遥感中复杂任务的自动化程度得以提高，涵盖变化检测、标注和多模态推理等领域 [3], [4] 。尽管如此，仍存在若干关键挑战：

(1) 缺乏环境上下文，因为大多数 VLM 仅依赖图像，忽略了对理解和观察地理变化至关重要的传感器数据 [3] ；(2) 单一来源注释偏差，使用单一语言模型进行数据注释限制了语言多样性并引入了偏差 [5] ；(3) 空间和语义多样性有限，因为广泛使用的数据集如 LEVIR-CD [6] 和 RSICap [7] 缺乏健全建模所需的地理范围和对象类别多样性 [8] ；以及 (4) 缺乏互动推理，当前的 VLM 没有提供支持规划和假设决策的“假如”界面 [3] 。

为了解决这些挑战，我们引入了 ChatENV，一个将遥感图像和真实环境传感器数据相结合的统一框架。这种结合提高了检测和解释随时间和空间变化的能力。我们方法的核心是一个大规模的卫星-航空数据集，该数据集覆盖了广泛的地理区域，包含多样的土地类型和物体类别，并将每组/对图像与真实的传感器读数（温度、湿度、风速、紫外线指数、PM_{10} 、CO 等）配对。我们的主要贡献是：

- 最先进的多模态数据集。我们发布了同类中最大的一个数据集，包含超过 177k 的卫星图像，分为 152k 的时间分割图像对，涵盖 62 个对象类别和 197 个国家。每对图像都有由双模型管道（GPT-4o 和 Gemini 2.0）生成的丰富变化描述，增加了语言多样性、上下文丰富性和改进的泛化能力。所有测试描述都经过人工审核以确保准确性，为环境变化检测任务提供了一个强有力的基准。
- 传感器引导的多模态学习。通过融合高分辨率图像与温度、湿度、风速、紫外线指数和排放数据（PM_{10} 、CO、NO_2 ），我们的模型不仅限于视觉变化，还包括环境上下文，从而更深入地理解变化的根本原因。
- 互动式视觉语言聊天模型 ChatENV 是一个多任务视觉语言对话模型，可以描述单张图像，识别并解释图像对之间的变化，并回答基于情景的问题（例如，“如果种植更多树木会怎样?”）。这支持探索现实和想象中的变化，以便进行环境知情的决策。

II. 相关工作

基准数据集: 有几个数据集支持地理空间视觉-语言任务。对于描述任务，RSICD [8] 、NWPU-Captions [9] 和 RSICap [7] 提供不同尺度的图像-文本对。RSVQA [10] 和基于 BigEarthNet 的扩展 [11] 专注于空间视觉问答 (VQA)，而 LEVIR-CD [12] 和 LEVIR-CC [13] 提供变化检测 (CD) 标签和自然语言描述。ChangeCLIP [14] 评估语义变化检测，而 SpaceNet [15] 提供长期城市影像，但缺乏文本注释。然而，大多数现有数据集缺乏时间的多样性、环境传感器数据或支持开放式用户查询。仍然需要支持语境理解、多模态学习和交互式推理的基准。

视觉 - 语言模型：最近的视觉语言模型（VLMs），如 Pixtral [16] 、DeepSeek-VL [17] 、Janus Pro [18] 、Phi-4-Multimodal [19] 和 LLaMA-3-Vision [20] ，在任务如图像标题生成、视觉推理和指令跟随上展示了强大的性能。这些模型直接将视觉编码器整合到大型语言模型（LLMs）中，并在大规模图像-文本数据集上进行训练。虽然这些模型并非为遥感设计，但其可扩展性和推理能力使其成为进行领域适应的有前途的候选者。

遥感领域的 VLMs：为了解决地理空间任务，已经出现了针对特定领域的 VLMs 改进版本。例如，RSGPT [7] 在 RSICap 上对 Vicuna 进行微调以生成空中描述。RS-LLaVA [21] 和 EarthGPT [22], [23] 使用基于 LLaMA 的 LLMs 结合 CLIP/ViT 骨干网络进行描述生成和视觉问答，其中整合了多传感器数据。RemoteCLIP 和 GeoChat [24], [25] 构建于大型的图像-文本语料库之上，但缺乏时间建模。GeoLLaVA [26] 使用基于视频的微调进行变化检测，尽管如此，它仍然缺乏注释多样性和交互性。

航拍图像中的字幕生成和视觉问答：遥感视觉问答的初步工作始于 RSVQA [10] ，使用 OpenStreetMap 进行空

All authors are with the Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE. Correspondence: hosam.elgendy@mbzuai.ac.ae.

[1]注释数据和代码可以通过以下网址在 github 上获取: https://github.com/HosamGen/ChatENV/

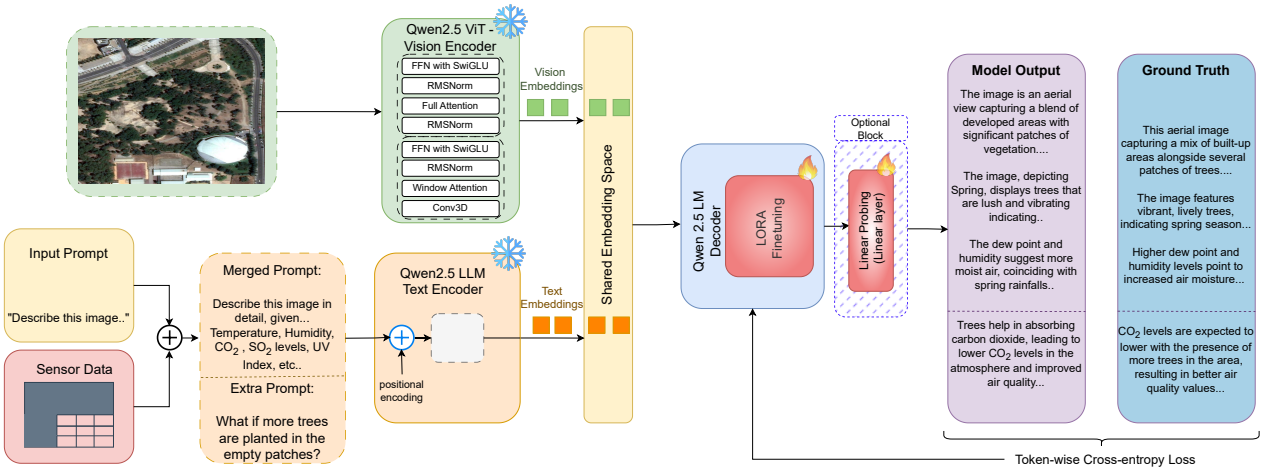

Fig. 1: ChatENV 的流程概述。航空 RGB 图块和传感器标记的提示（例如，温度、湿度、CO_2 ）分别通过冻结的 Qwen 2.5 ViT 和文本编码器进行编码。它们的嵌入被投影到共享空间中以对 Qwen 2.5 解码器进行条件化，仅训练 LoRA 适配器和一个可选的线性探测器。通过丰富的字幕上的词级交叉熵来训练模型以 (i) 描述场景，(ii) 推理当前的环境数据，以及 (iii) 回答“如果”类型的查询。

TABLE I: 用于遥感任务的视觉语言模型比较。基于传感器输入、会话和时间能力以及预测能力。

Model	Interactive	Temporal	What-if	Sensor
RSGPT [7]	✗	✗	✗	✗
RS-LLaVA [22]	✓	✗	✗	✗
EarthGPT [23]	✓	✗	✗	✗
RemoteCLIP [24]	✗	✗	✗	✗
GeoChat [25]	✓	✗	✗	✗
GeoLLaVA [26]	✓	✓	✗	✗
Tree-GPT [28]	✓	✗	✗	✗
ChangeCLIP [14]	✗	✓	✗	✗
CD-VQA [11]	✗	✓	✗	✗
TEOChat [29]	✓	✓	✓	✗
Ours: ChatENV	✓	✓	✓	✓

间问题，并由 [11] 扩展至 BigEarthNet。这些工作依赖于合成提示，并缺乏环境数据。在字幕生成方面，RSICD 引入了 CNN-RNN 管道，NWPU-Captions 加入了注意力机制，最近的工作 [27] 探索了元数据意识的预训练，但仍局限于静态图像。

交互式推理用于环境监测：像 RS-LLaVA 和 EarthGPT 这样的对话视觉语言模型 (VLM) 能够进行多轮问答，但缺乏预测推理。Tree-GPT [28] 添加了用于基于命令的交互的地理空间工具，但不能模拟假想的、反事实的场景。最后，TEOChat [29] 通过询问关于图像中变化的物体或影响某个地点的灾害的问题来支持场景监测，但它没有结合基于温度或其他气象指标的环境推理。

大多数之前的工作要么 (i) 仅处理静态图像（缺乏时间信息），要么 (ii) 省略连续传感器信号，或者 (iii) 使用模板化查询，或 (iv) 缺乏反事实推理。ChatENV 通过整合时间图片对和传感器数据以进行交互和预测的环境理解，解决了以上所有四个问题，如表 I 所述。

III. 数据收集和预处理

我们构建了一个多步骤流程，用于从 fMoW 数据集中获取和预处理高分辨率卫星图像，概述见图 2 。

从 fMoW 数据集中构建时间对。fMoW RGB 数据集 [30] 包含 363,571 张训练图像和 53,041 张验证图像，全球覆盖年份从 2002 到 2017 。空间分辨率约为 0.3 m ，它提供了在 62 类别之间进行细粒度变化检测的高精度 [31] 。我们利用数据集的元数据根据位置配对图像，同时保持至少 12 个月的时间间隔，因为初步检查表明场景中的可见和可检测变化通常在大约一年后出现。经过筛选和去重后，最终训练集包含 $177k$ 张图像，相当于 $152k$ 对图像，而测试集包含 $14k$ 张图像，相当于 $10k$ 对图像。表 II 和图 4 分别提供了 62 个类别在 197 个国家的概况和视觉分布。

我们使用了 fMoW 数据集的元数据，该元数据包含地理位置、国家代码、时区和时间戳，通过第三方 API 获取天气和大气排放数据。对于天气特征，VisualCrossing API 使用每个图像的拍摄位置和时间戳获取了一大组字段，如 `temperature` 、`dew point` 、`humidity` 等。同样，排放数据是通过 Open-Meteo 为欧洲样本和 OpenAQ 为非欧洲样本收集的，生成的参数如 $pm2.5$ 、CO 、NO_2 和 `ozone` 。我们专注于出现频率最高且分析最相关的特征，并对 fMoW 数据集中的国家代码与 API 返回的国家名称进行了交叉参考，以确保时间上的对齐和特定季节的背景。与有限相关性或大量缺失数据的特征被丢弃，而缺乏排放数据的特征被保留以保持类别和地理多样性。这种环境元数据与数据集全球多年度覆盖的融合映射使我们能够捕捉时间、气候和大气变化，同时最大程度地减少信息缺失，以支持下游建模工作。

a) 注释生成： 我们制作了描述性注释以捕捉每个图像的最重要特征，并总结匹配样本之间的时间变化。为了扩展泛化、减轻风格偏见和增强稳健性，我们使用了两个最新的 LLM：GPT-4o 和 Gemini 2.0 Flash Thinking。所有提示都包含季节、语义类别和环境传感器数据，并针对每个模型采用一致的提示策略。数据集在 ChatGPT 和 Gemini 之间平均分配，来自同一位置的图像分配给同一个模型，以实施一致的局部划分。训练集包含 177 ， 719 张图像（152 ， 710 对），测试集包含 14 ， 806 张图像（10 ， 020 对），每个图像都增加了天气数据和文本注释。在这些图像中，93 ， 834 张训练图像和 14 ， 322 张验证

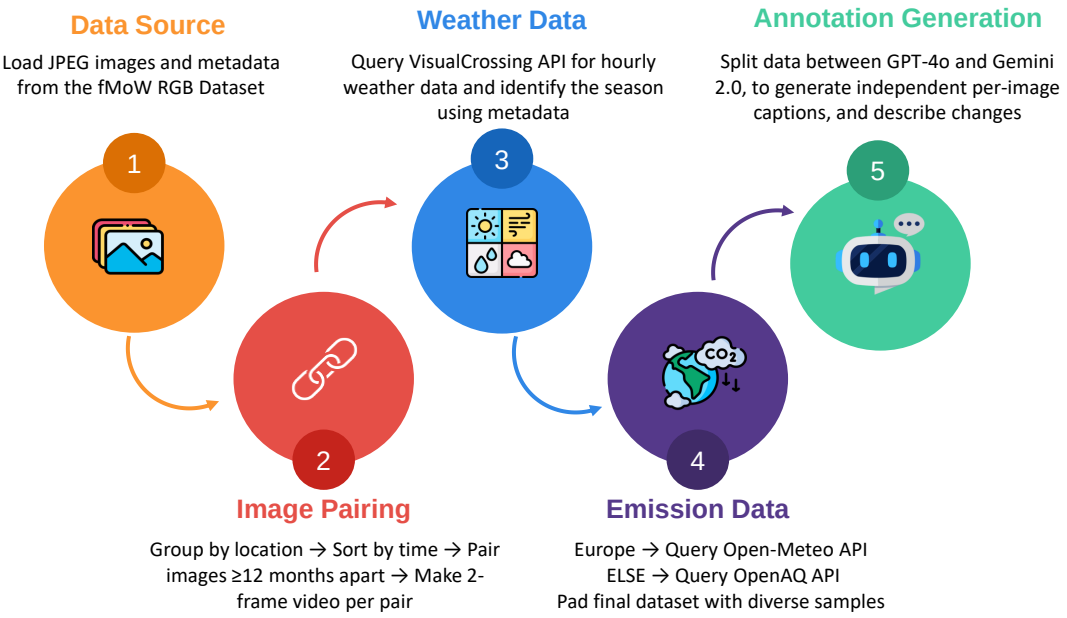

Fig. 2: 用于环境变化分析的预处理流程的视觉概览。该过程从卫星图像获取（数据来源）开始，接着进行时间配对（图像配对），整合气象数据，结合排放数据（排放数据）进行丰富，并通过 GPT-4o 和 Gemini 2.0 进行注释生成（注释生成）。

TABLE II: 原始 fMoW 数据集与包含环境数据的修改版本之间的比较。

	# Images (train/test)	# Annotated Pairs (train/test)	Weather Sensors	Emissions Sensors
fMoW Dataset	363,571 / 53,041	0	None	None
Modified Dataset	177,719 / 14,806	152,710 / 10,020	Temp, Humidity, Dew, etc.	CO, O_3, NO_2, etc.

图像还包含排放数据，从而能够对大气因素进行端到端分析。

在进行模型评估之前，测试数据集经过了彻底的人工检查，以确保 GPT-4 和 Gemini 2.0 模型生成的标注的可靠性和正确性。首先，标注者（他们也是作者）验证了每张图像中的主要对象是否与原始数据集中提供的真实类别标签一致，这为标注质量提供了初步的验证层。其次，针对每个样本，标注者被问及特定问题，以评估生成的描述对于图像中关键视觉元素和显著对象的准确性。每位标注者被要求根据三个问题，对每个描述以 1 到 5 的评分进行评价。对每个描述的评分进行汇总，并使用 9 分作为阈值来筛选进入测试集的标注与未进入的标注，具体如图 3 所述。为了考虑到评估此类开放性描述时固有的主观性，数据集在标注者之间进行了不重叠的分配，从而避免了评估者之间的差异。

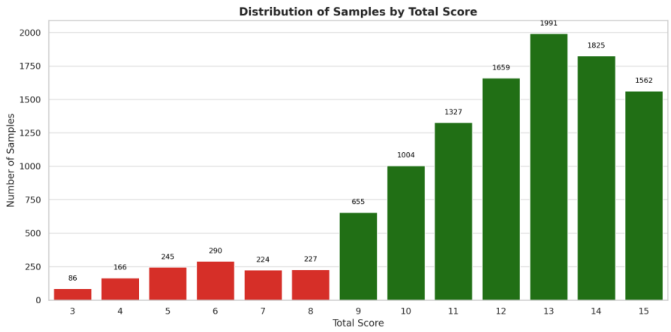

Fig. 3: 通过对测试集进行人工评估，按分数对样本进行总分布。每个样本在三个标准上被评为 1-5 分。总评分超过 9 分的样本被保留为测试集。

IV. 方法论

我们的方法利用了两种技术，一种基于会话 VLMs，另一种基于视频 VLMs，每种技术都针对我们任务所需的不同功能进行了优化。会话 VLMs 经过微调，用于互动对话，记忆图像特征，从而支持用户提出的动态比较和“假设”情况，使其具有交互性。另一方面，视频 VLMs 被优化用于跟踪一对图像之间的时间变化，这有助于详细描述每个单独图像及其变化。我们利用视频 VLMs 来对它们进行时间理解比较，并与我们的会话 VLM ChatENV 进行比较。

Qwen VLM：我们在 Qwen2.5-VL-7B-Instruct 模型上训练 ChatENV，该模型结合了一个 Vision Transformer (ViT) 图像编码器与一个大语言模型解码器 [32] 。这个模型是 Qwen 2.5 模型的多模态版本，是一个开源、轻量级的模型，易于微调，并在 VQA 和推理基准测试上表现出色。如图 1 所示，ViT、文本编码器和投影层保持冻结状态。微调仅限于放置在解码器内每个注意力和前馈块内的轻量级低秩适应（LoRA）适配器。可选地添加一个线性探针对标量预测任务进行。一个训练样本包括一个 RGB 航拍图块和一个增强了结构化传感器数据的用户提示。由于无法直接从图像中推断环境传感器读数（例如，温度、一氧化碳），我们将其明确地作为用户输入提供。

ChatENV 可以支持三种互为补充的任务。在单轮情况下，每个样本都有一个图像和一个指令，例如：“描述这张图片”，模型必须生成一个单一的真实标签。在双轮假设情况下，模型对图像进行描述，然后回答一个假设性的“如果 ... 会发生什么”问题。在三轮区别情况下，模型顺序对

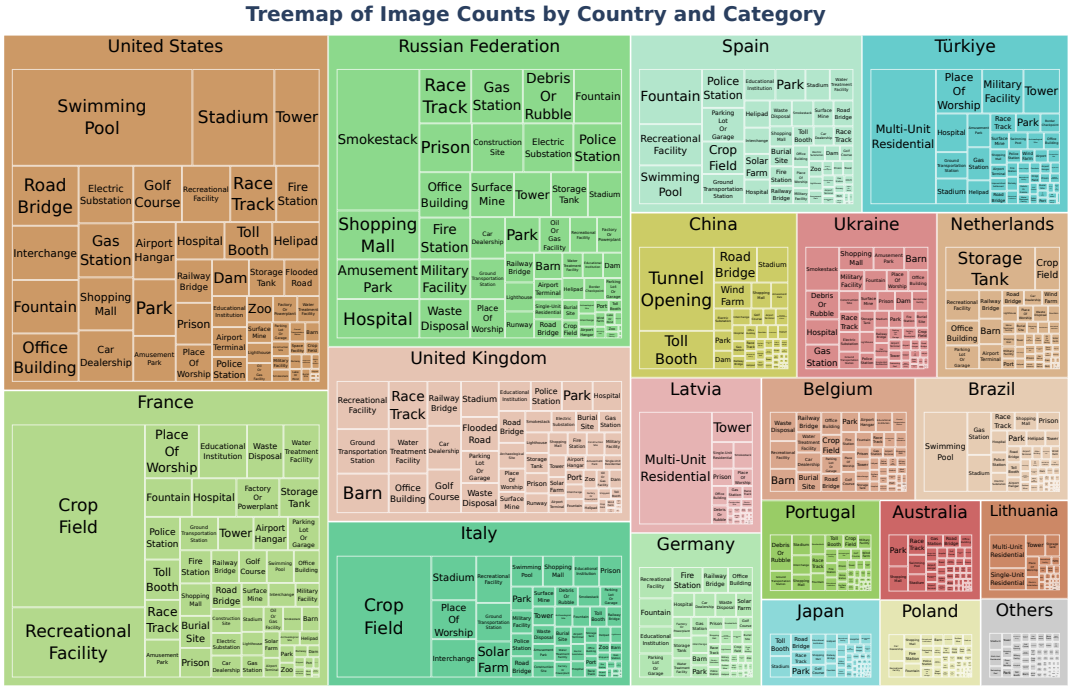

Fig. 4: 树状图可视化显示按国家和类别划分的卫星图像数量分布。较大的矩形表示数据集中更频繁出现的物体类别，如游泳池、体育场和道路。国家间的空间多样性，如法国和意大利的农田，或美国和俄罗斯联邦的城市结构，突显了广泛的地理和语义覆盖，这对于稳健的变化分析至关重要。

两个图像进行描述，然后回答“有什么区别?”的提示。我们在纯描述、比较推理和假设任务中进行均匀采样，并在每一个助手轮次上训练交叉熵损失。

基于视频的 VLMs：我们采用了两个最新的 SOTA 模型：LLaVA-NeXT-Video [33] 和 Video-LLaVA [34]。LLaVA-NeXT-Video 是 LLaVA-NeXT 的一个视频适配版本，利用先进的视觉推理和改进的通用知识，提供准确的帧级描述和可靠的时间比较，特别是在零样本视频理解场景中表现出色。同样，Video-LLaVA 有效地将时间视觉信息与语言理解相结合，在视频语言基准测试中实现优越性能，使其非常适合表达细微的视觉变化。

在微调期间，我们将每对图像转换为一个 2 帧的视频序列，其中每个图像都是一个单独的帧。通过视频编码器对帧进行均匀采样以创建视觉嵌入，同时对文本提示进行编码。视觉和文本特征在训练期间被优化以实现对齐，从而增强这些模型分别描述每张图像并准确找到精细差异的能力。我们选择了 7B 参数的模型变体，以便在一块 GPU 上运行，出于实际考虑。

V. 实验设置

LoRA 我们使用 LoRA [35] 微调了我们的模型。LoRA 认为预训练和微调的权重之间存在低秩差异。给定一个预训练的矩阵 $W_0 \in \mathbb{R}^{d\times k}$，LoRA 将更新分解为 $W_0 + \Delta W = W_0 + BA$，其中 $B \in \mathbb{R}^{d\times r}$、$A \in \mathbb{R}^{r\times k}$ 和 $r \ll \min(d,k)$。这里，W_0 保持不变，而 A 和 B 是可训练的。前向传播变为 $h = W_0 x + BAx$，由 α/r 缩放，α 控制更新幅度。通常，A 是随机初始化的，而 B 是零初始化的，确保 ΔW 从零开始。

评估指标：我们采用五种指标来严格评估生成质量，这些指标能够捕捉语义相似性和表面重叠，并为参考文本对齐提供补充性见解。

(I) ROUGE [36] 计算生成文本和参考文本之间的 n - gram 召回。我们报告 ROUGE-L，它奖励最长公共子序列 (LCS)，因而反映流畅性和短语级别的重叠。ROUGE-1 和 2 评估一元和二元组召回，反映词汇重叠。ROUGE 特别适合用于内容覆盖率估计。

(II) SBERT 相似性使用 Sentence-BERT 将生成的文本和参考文本编码为高维向量，并使用余弦相似度 [37] 进行比较。即使在重新措辞的句子中，它也能检测到语义重叠。

(III) BERTScore [38] 通过使用上下文嵌入对生成和参考标记进行比较计算得出。我们使用 DeBERTa V3 来计算准确率、召回率和 F1 分数，这些指标反映了不同措辞之间的流畅性和语义相似性。

(IV) COMET [39] 是一种神经度量，它经过了人类判断的训练。它将源文本、假设和参考投射到一个公共空间，并对充分性和流畅性的质量评分进行预测。COMET 对生成任务特别稳健，并且与人类评估具有很强的相关性。

(V) 关键词聚类评估 (KCE) 是我们引入的一个指标，用于评估方向性语言。诸如上升、增加、下降和减少等词语对于比较性描述是必不可少的，但在标准指标中它们常常被忽视或处理不当。我们从“差异”句子中提取常见的方向性术语，将它们的变体分组成簇，并基于它们在参考和假设中的存在情况计算 F1 分数 (KCE-F1)。这反映了模型检测和保持意义中感知变化的能力。

真实标注包括：(i) 图像描述，(ii) 图像差异描述，(iii) 基于差异的假设问题，以及 (iv) 基于第二幅图像的回答。例子：如果第一幅图像有较少的建筑物、增加的风速和升

TABLE III: 在 ChatENV 上的模型性能（基于 Qwen），每列的最佳用下划线标出，每组的最佳用颜色标出。

(a) 三折设定

Annotations	Training					
ChatGPT	Base	0.124	0.430	0.824	0.496	0.710
	LoRA	0.242	0.702	0.889	0.733	0.818
	Lin. PROBE	0.233	0.648	0.884	0.699	0.817
Gemini	Base	0.122	0.450	0.825	0.490	0.692
	LoRA	0.298	0.803	0.902	0.763	0.826
	Lin. PROBE	0.289	0.794	0.899	0.752	0.830
ChatGPT + Gemini	Base	0.124	0.445	0.825	0.495	0.705
	LoRA	0.250	0.737	0.890	0.745	0.814
	Lin. PROBE	0.236	0.706	0.883	0.713	0.809

(b) 两回合（假设）设置

Training					
Base	0.108	0.607	0.840	0.627	0.587
LoRA	0.231	0.597	0.893	0.686	0.800
Lin. PROBE	0.233	0.596	0.889	0.684	0.813
Base	0.115	0.597	0.837	0.626	0.569
LoRA	0.282	0.667	0.900	0.705	0.816
Lin. PROBE	0.247	0.647	0.889	0.702	0.759
Base	0.111	0.603	0.838	0.628	0.576
LoRA	0.254	0.646	0.895	0.695	0.809
Lin. PROBE	0.234	0.629	0.886	0.666	0.785

高的温度，则一个示例问题可能是："如果出现更多的建筑物会怎样?" 一个示例参考答案可能是："由于建筑物诱发的阻力，风速可能减缓，由于更多的混凝土表面，温度可能上升。" 在假设推理过程中，模型仅根据第一幅图像和假设情境生成答案，而不查看第二幅图像。模型输出使用相同的测试数据对每种配置进行参考答案测试。

所有模型都在 48GB A6000 GPU 上对整个数据集进行了一轮训练。对于 ChatGPT 和 Gemini 特定的注释进行了单独的进一步测试。对于基于视频的 VLMs，单独为 ChatGPT 和 Gemini 构建了一个大约占数据集 10 % 的样本集，并用于测试数据大小对模型性能的影响。

VI. 结果 & 讨论

三轮评估。表 IIIa 展示了 ChatENV 在一个三轮任务上的表现，该任务 (i) 描述图像 1，(ii) 描述图像 2，以及 (iii) 解释它们的区别。训练分为三种注释模式：ChatGPT、Gemini（各涵盖 50 % 的数据），以及它们的联合体，评估则是在一个由两个来源的注释组成的固定测试集上进行的。此外，传感器信息作为输入提示的一部分传递给模型中的每对图像。所有模型共享相同的 Qwen-2.5 骨干和解码器超参数；唯一的区别在于适应策略。

零样本参考：通过直接传递真实标注而不进行任何微调来评估基础模型，作为零样本基线。正如预期的那样，由于基础模型在预训练期间接触过卫星影像，其在大多数指标上的表现尚可。然而，微调在所有指标上均带来了一致的提升。

LoRA 与线性探测：持续地植入 LoRA 适配器可在所有测量中提升相对于基线的表现，在单独和组合注释条件下均如此。重要的是，对于组合集，LoRA 将 SBERT 和 COMET 的分数提高了约 +0.30 。然而，线性探测在与 LoRA 相比时具有竞争力且指标接近，并且可能在 KCE-F1 分数上超过 LoRA。查看各个集合，LoRA 在所有五个指标中都优于探测方法，并在仅限 ChatGPT 的条件下达到最佳分数。这表明 LoRA 提升了语义匹配和上下文理解，而在线性探测作为一种有效的轻量选项时，当表面相似性最重要或资源有限时，它显得更为适用。

两轮 "如果" 评估。表 IIIb 显示了当模型输入单个图像和传感器数据时，假设分析 的表现，生成一个天气感知的标题，然后通过预测环境变化如温度上升或 CO_2 减少来回应一个假设性问题（例如，"如果种植更多的树会怎样?"）。得分通常与表 IIIa 中看到的趋势一致，但在某些指标上通常较低，因为模型必须 假设 情况，因为没有第二个图像，使任务本质上更困难。尽管复杂性增加，KEC-F1 和 SBERT 分数仍然很高，这验证了模型准确识别重要变化的方向和意义，这是假设情景的总体目标。

与最先进的视频视觉语言模型的比较。表 IV 展示了一项消融研究，比较了 ChatENV 与两个顶尖的视频模型，Video-LLaVA 和 LLaVA-NeXT-Video 架构，这些架构专为时间理解而设计。这两个模型将两个输入帧视为一个简短的剪辑，并在一个三轮设置上进行训练，使它们能够直接记忆。这在大多数汇总指标上实现了适度的优势；例如，最佳视频配置的 COMET 为 0.833，而 ChatENV 为 0.763。

最重要的是，ChatENV 在某些指标上仍能表现出色并超过视频模型。它实现了最佳的 BERT-F1（0.902 对比 0.818）和 KCE-F1（0.830 对比 0.758）分数，并且几乎相同的 ROUGE-L 得分，这反映了尽管在缺乏连续视频流的情况下需要推断时间变化，仍能更好地与人工标注数据保持一致。

视频模型的基础（未经调优的）版本表现不如 ChatENV 基线，强调了领域曝光所起的作用：Qwen 的预训练语料库中已经包含卫星数据，而 Video-LLaVA 和 LLaVA-NeXT 没有航空数据。在使用我们精心策划的数据集进行微调后，视频模型的性能至少提高了 75 % ，这表明建议的数据处理流程是如何克服这种领域差距的。

从表 IIIa 和 IV 可以明显看出，Gemini 注释在几乎所有指标上始终优于 ChatGPT 注释。从表 IV 进一步比较这两个模型可以看出，使用更少的样本时，Video-LLaVA 在 ROUGE-L、COMET 和 KCE-F1 等指标上往往优于 LLaVA-NeXT-Video，其中 Video-LLaVA 连同 Gemini 注释一起，获得了最佳的总体得分。

表格 **??** （左）展示了提示中包含传感器信息与不包含传感器信息的效果。此实验是在 ChatENV 模型中使用三回合设置和 LoRA 微调进行的。很明显，对于所有的注释集，包含传感器数据在所有指标中都获得了更高的分数，这表明传感器数据对于进一步解释和描述图像中的某些场景是必要的。此外，我们将我们的模型与 SOTA 变更检测模型进行了比较，包括 TEOChat 和 RS-LLaVA。虽然 TEOChat 可以同时处理两张图像，但它在描述场景层次上的变化时常常困难重重，表现较好的是当具体物体被强调或当问题集中于影响该地区的自然灾害时。另一方面，RS-LLaVA 处理单张图像，类似于 ChatENV，但在将传感器读数与对应的地理环境联系起来时存在困难。相比之下，ChatENV 通过提供更全面的场景层次的变化描述，并将传感器数据连贯地整合到其响应中，从而优于这两个模型。

图 5 提供了 What-If 交互流程的定性结果。用户首先向

TABLE IV: Video-LLaVA（左）和 LLaVA-NeXT-Video（右）的比较。每个模型每个指标的最佳结果用色彩标注。所有实验中每个指标的最佳结果用下划线表示。

Annotations	Training	ROUGE-L		SBERT		BERT-F1		COMET		KCE-F1	
ChatGPT	Base	0.116	0.118	0.562	0.563	0.540	0.546	0.436	0.444	0.495	0.500
	10K	0.255	0.172	0.826	0.696	0.757	0.674	0.785	0.641	0.709	0.649
	76K	0.224	0.243	0.859	0.868	0.743	0.776	0.789	0.800	0.736	0.758
	Lin. Probe	0.147	0.206	0.675	0.767	0.634	0.731	0.647	0.761	0.622	0.652
Gemini	Base	0.073	0.079	0.551	0.552	0.512	0.519	0.414	0.421	0.450	0.455
	10K	0.299	0.243	0.897	0.896	0.818	0.804	0.833	0.831	0.729	0.717
	76K	0.246	0.238	0.868	0.894	0.774	0.770	0.805	0.824	0.720	0.732
	Lin. Probe	0.194	0.242	0.783	0.819	0.694	0.754	0.716	0.768	0.622	0.636
ChatGPT+Gemini	Base	0.095	0.099	0.557	0.558	0.526	0.553	0.425	0.433	0.472	0.478
	15K	0.238	0.208	0.861	0.820	0.770	0.755	0.810	0.806	0.639	0.688
	152K	0.234	0.205	0.849	0.809	0.784	0.769	0.816	0.812	0.757	0.704
	Lin. Probe	0.215	0.201	0.825	0.705	0.735	0.761	0.748	0.798	0.626	0.685

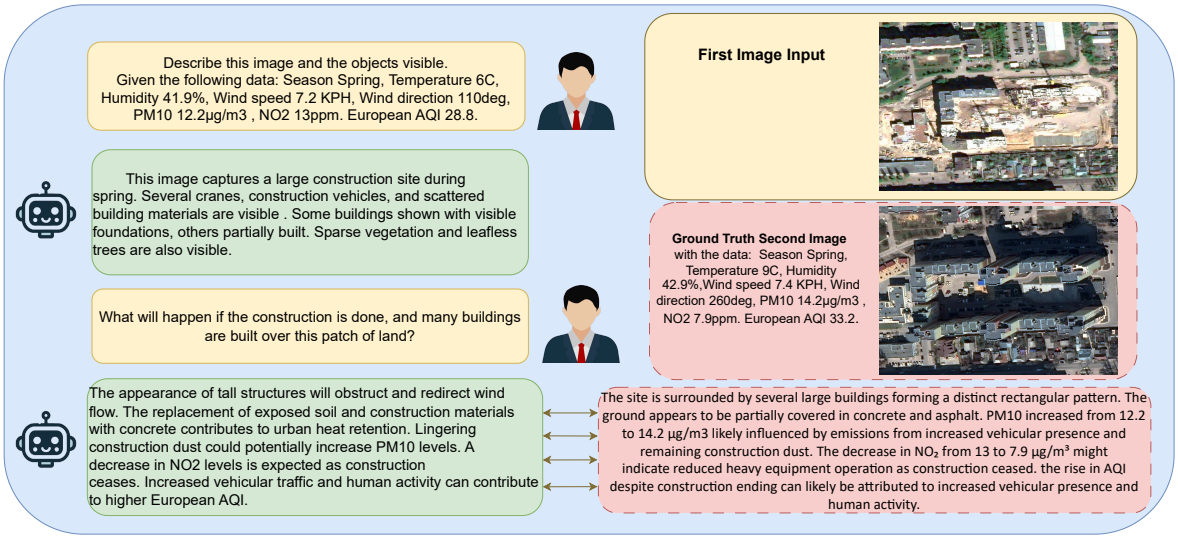

Fig. 5: 该图展示了与 ChatENV 的假设性交互。在给定初始图像和环境元数据的情况下，用户提出了一个场景问题："如果这块土地上进行施工并建造许多建筑物，会发生什么?" 模型生成了一个详细的回答，与第二张（真实情况）图像的描述非常匹配，例如由于灰尘和交通导致的 PM10 增加，施工后 NO_2 减少，风向模式改变，以及混凝土造成的热量滞留。

模型查询带有其关联传感器读数的航空图像，然后提出一个假设性（"如果 ...会怎么样呢?"）问题。ChatENV 的回答（绿色）与根据第二幅图像（红色）得出的参考答案非常接近，在这个情境中，What-if 情景答案实际上发生（使用数据集中的第二幅图像）。这表明其有能力推理关于未见过的结果。这种准确性支持实际用例，例如环境监测和城市规划模拟，在这些情况下，可靠但富有想象力的预测至关重要。

本文介绍了 ChatENV，这是一种新型的互动式环境 VLM，能够联合推理卫星图像和真实世界的传感器数据。通过创建同类中最大的多模态数据集，结合航空图像和丰富的环境元数据对，以及异构的双模型注解，ChatENV 实现了全面且具上下文意识的环境理解。该模型促进了场景的细粒度描述、时空比较和 "假如" 情景模拟。实验结果表明，ChatENV 能够输出有意义且人性化的描述，这对于城市规划和环境监测非常重要。

未来工作：我们旨在 (i) 整合更多的信息源（例如，SAR、多光谱、社会经济数据），(ii) 引入记忆增强型代理以进行长期推理，(iii) 通过持续的 LoRA 更新来支持实时部署，(iv) 探索联合视频-对话的预训练。这些步骤将 ChatENV 定位为预测环境智能的基础模型。

References

[1] M. A. Wulder, J. G. Masek, W. B. Cohen, T. R. Loveland, and C. E. Woodcock, "Landsat continuity: Issues and opportunities for land cover monitoring," *Remote Sensing of Environment*, vol. 122, pp. 84–91, 2012.

[2] C. Corradino, G. Ganci, A. Cappello, G. Bilotta, A. Hérault, and C. Del Negro, "Mapping recent lava flows at mount etna using multispectral sentinel-2 images and machine learning techniques," *Remote Sensing*, vol. 11, no. 16, p. 1916, 2019.

[3] C. Liu, K. Chen, H. Zhang, Z. Qi, Z. Zou, and Z. Shi, "Change-agent: Towards interactive comprehensive remote sensing change interpretation and analysis," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[4] Z. Zhang, H. Shen, T. Zhao, B. Chen, Z. Guan, Y. Wang, X. Jia, Y. Cai, Y. Shang, and J. Yin, "Georsmllm: A multimodal large language model for vision-language tasks in geoscience and remote sensing," 2025. [Online]. Available: https://arxiv.org/abs/2503.12490

[5] Z. Wang, R. Prabha, T. Huang, J. Wu, and R. Rajagopal, "Skyscript: A large and semantically diverse vision-language dataset for remote sensing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5805–5813.

[6] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sensing*, vol. 12, no. 10, p. 1662, 2020.

[7] Y. Hu, J. Yuan, C. Wen, X. Lu, Y. Liu, and X. Li, "Rsgpt: A remote sensing vision language model and benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 224, pp. 272–286, 2025.

[8] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, 2017.

[9] Q. Cheng, H. Huang, Y. Xu, Y. Zhou, H. Li, and Z. Wang, "Nwpu-captions dataset and mlca-net for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[10] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "Rsvqa: Visual question answering for remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 12, pp. 10 129–10 141, 2021.

[11] Z. Yuan, L. Mou, Z. Xiong, and X. X. Zhu, "Change detection meets visual question answering," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[12] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sensing*, vol. 12, no. 10, p. 1662, 2020.

[13] C. Liu, R. Zhao, H. Chen, Z. Zou, and Z. Shi, "Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2022.

[14] S. Dong, L. Wang, B. Du, and M. Meng, "Changeclip: Remote sensing change detection with multimodal vision-language representation learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 208, pp. 53–69, 2024.

[15] A. Van Etten, D. Hogan, J. M. Manso, J. Shermeyer, N. Weir, and R. Lewis, "The multi-temporal urban development spacenet dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6398–6407.

[16] P. Agrawal, S. Antoniak, E. B. Hanna, B. Bout, D. Chaplot, J. Chudnovsky, D. Costa, B. D. Monicault, S. Garg, T. Gervet, S. Ghosh, A. Héliou, P. Jacob, A. Q. Jiang, K. Khandelwal, T. Lacroix, G. Lample, D. L. Casas, T. Lavril, T. L. Scao, A. Lo, W. Marshall, L. Martin, A. Mensch, P. Muddireddy, V. Nemychnikova, M. Pellat, P. V. Platen, N. Raghuraman, B. Rozière, A. Sablayrolles, L. Saulnier, R. Sauvestre, W. Shang, R. Soletskyi, L. Stewart, P. Stock, J. Studnia, S. Subramanian, S. Vaze, T. Wang, and S. Yang, "Pixtral 12b," 2024. [Online]. Available: https://arxiv.org/abs/2410.07073

[17] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang, Y. Sun, C. Deng, H. Xu, Z. Xie, and C. Ruan, "Deepseek-vl: Towards real-world vision-language understanding," 2024. [Online]. Available: https://arxiv.org/abs/2403.05525

[18] X. Chen, Z. Wu, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, and C. Ruan, "Janus-pro: Unified multimodal understanding and generation with data and model scaling," *arXiv preprint arXiv:2501.17811*, 2025.

[19] A. Abouelenin, A. Ashfaq, A. Atkinson, H. Awadalla, N. Bach, J. Bao, A. Benhaim, M. Cai, V. Chaudhary, C. Chen *et al.*, "Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras," *arXiv preprint arXiv:2503.01743*, 2025.

[20] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[21] Y. Bazi, L. Bashmal, M. M. Al Rahhal, R. Ricci, and F. Melgani, "Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery," *Remote Sensing*, vol. 16, no. 9, p. 1477, 2024.

[22] Y. Bazi, L. Bashmal, M. M. Al Rahhal, E. Ricci, and F. Melgani, "Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery," *Remote Sensing*, vol. 16, no. 9, p. 1477, 2024.

[23] W. Zhang, M. Cai, T. Zhang, Y. Zhuang, and X. Mao, "Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–11, 2024.

[24] F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, Q. Ye, L. Fu, and J. Zhou, "Remoteclip: A vision language foundation model for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024. [Online]. Available: https://doi.org/10.1109/TGRS.2024.3390838

[25] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, "Geochat: Grounded large vision-language model for remote sensing," *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[26] H. Elgendy, A. Sharshar, A. Aboeitta, Y. Ashraf, and M. Guizani, "Geollava: Efficient fine-tuned vision-language models for temporal change detection in remote sensing," *arXiv preprint arXiv:2410.19552*, 2024. [Online]. Available: https://arxiv.org/abs/2410.19552

[27] J. Bourcier, G. Dashyan, K. Alahari, and J. Chanussot, "Learning representations of satellite images from metadata supervision," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[28] S. Du, S. Tang, W. Wang, X. Li, and R. Guo, "Tree-gpt: Modular large language model expert system for forest remote sensing image understanding and interactive analysis," *arXiv preprint arXiv:2310.04698*, 2023.

[29] J. A. Irvin, E. R. Liu, J. C. Chen, I. Dormoy, J. Kim, S. Khanna, Z. Zheng, and S. Ermon, "Teochat: A large vision-language assistant for temporal earth observation data," *arXiv preprint arXiv:2410.06234*, 2024.

[30] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, "Functional map of the world," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6172–6180.

[31] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. Lobell, and S. Ermon, "Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery," *Advances in Neural Information Processing Systems*, vol. 35, pp. 197–211, 2022.

[32] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," 2025. [Online]. Available: https://arxiv.org/abs/2502.13923

[33] H. Liu, Z. Wu, C. Li, J. Yang, L. Li, Z. Huang, and J. Gao, "Llava-next: Next-generation large vision-language models with decoupled multimodal pre-training," 2024.

[34] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan, "Video-LLaVA: Learning united visual representation by alignment before projection," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 5971–5984. [Online]. Available: https://aclanthology.org/2024.emnlp-main.342/

[35] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[36] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[37] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 3982–3992.

[38] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020.

[39] R. Rei, A. Farinha, A. Lavie, and A. F. T. Martins, "Comet: A neural framework for mt evaluation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 2685–2702.

Hosam Elgendy is currently a Research Associate in the Machine Learning department at Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) in Abu Dhabi, UAE. He previously received his Master of Science degree in Computer Vision from MBZUAI in 2024. He obtained his Bachelor of Science in Electrical Engineering from the American University of Sharjah (AUS), Sharjah, United Arab Emirates. His research focuses on vision–language models and multi-modal learning systems. He worked extensively on integrating temporal reasoning, conversational AI, and sensor data into remote sensing applications. His interests span computer vision, natural language processing, and multi-modal data fusion, with a focus on scalable model evaluation and deployment.

Ahmed Sharshar is currently pursuing a PhD in Computer Vision at Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) in Abu Dhabi, UAE. He previously received his Master of Science degree in Computer Vision from MBZUAI. He obtained his Bachelor of Engineering degree in Computer Engineering from the Egypt-Japan University of Science and Technology (E-JUST), Egypt. His research primarily focuses on developing lightweight models and expanding their applications across various domains, such as natural language processing, computer vision, and human-computer interaction. Specifically, he aims to make these models more efficient and accessible, ensuring broader usability and practical deployment.

Ahmed Hesham is an AI researcher and engineer specializing in Natural Language Processing and speech technologies. He earned his Master' s degree in Natural Language Processing from the Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) in Abu Dhabi, UAE, and his Bachelor' s degree in Computer Engineering from the Egypt-Japan University of Science and Technology (E-JUST), Egypt. His research spans large language models, retrieval-augmented generation, and speech processing, with applications in healthcare, multilingual NLP, and Arabic language technologies. He has published in leading international venues such as INTERSPEECH, IEEE Access, and IJCNN. Ahmed' s work focuses on building efficient, inclusive, and production-ready AI systems, aiming to advance both cutting-edge research and real-world deployment.

Mohsen Guizani (S' 85, M'89, SM'99, F' 09) received his B.S. (with distinction) and M.S. degrees in electrical engineering, and M.S. and Ph.D. degrees in computer engineering from Syracuse University, New York, in 1984, 1986, 1987, and 1990, respectively. He is currently a Professor with the Machine Learning Department, Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE. Previously, he served in different academic and administrative positions at the University of Idaho, Western Michigan University, the University of West Florida, the University of Missouri-Kansas City, the University of Colorado-Boulder, and Syracuse University. His research interests include wireless communications and mobile computing, computer networks, mobile cloud computing, security, and smart grid. He was the Editor-in-Chief of IEEE Network. He serves on the Editorial Boards of several international technical journals, and is the Founder and Editor-in-Chief of the Wireless Communications and Mobile Computing journal (Wiley). He is the author of nine books and more than 500 publications in refereed journals and conferences. He has guest edited a number of Special Issues in IEEE journals and magazines. He has also served as a TPC member, Chair, and General Chair of a number of international conferences.

Total Score	Number of Samples
3	86
4	166
5	245
6	290
7	224
8	227
Subtotal (< 9)	1,238
9	655
10	1,004
11	1,327
12	1,659
13	1,991
14	1,825
15	1,562
Subtotal (≥ 9)	10,023
Total	11,261

TABLE V: 样本按总分数的分布。

Appendix A
注释细节和评分标准

这些是基于标准对 GPT-4 和 Gemini 2.0 注释进行人工评估的评分详细信息。这些信息在审查样本时呈现给注释者（作者）。请注意，由于评分样本时存在的主观性，整个集合在注释者之间进行了拆分，没有重叠。这消除了评审者之间的重叠。

A.

评分标准

问题 1：每个描述是否准确，并捕捉到图像中实际可见的细节?

- 1 点：该描述未捕捉图像中的任何细节。
- 2 分：描述捕捉到了一些细节，但遗漏了关键的可见元素/物体或包含不准确之处。
- 3 分：描述捕捉了大部分主要细节，但省略了次要但可见的元素/对象。
- 4 分：描述大部分准确，包括大多数关键细节，只有极少遗漏。
- 5 分：描述包括图像中可见的所有物体，并以高度准确性描述细节。

问题 2：描述是否正确包含或描述了图像的“类别”?

- 1 分：描述完全误识别或忽略了类别标签。
- 2 分：描述试图包括类别，但不正确或不完整。
- 3 分：描述包含类别，但表示部分或不清晰。
- 4 分：描述包括了类别，但可能缺乏上下文或支持细节。
- 5 分：描述准确地识别了类别，并在图像上下文中完全整合。

问题 3：模型在图像中更关注多数类还是少数类?

- 1 分: 描述未能集中于多数类或少数类。
- 2 分：该描述仅关注于多数类别，完全忽略少数类别。
- 3 分：描述优先考虑多数类，对少数类给予最少关注。
- 4 分：描述对多数类和少数类给予了均衡的关注，但细节不足。
- 5 分：描述对多数类和少数类给予同等和详细的关注。

基于这种评分，每个问题得到一个 3 到 15 之间的分数。我们选择保留得分在 9 或以上的样本，如表 V 所述。

TABLE VI: 在 LoRA 设置下，对 Video-LLaVA 和 LLaVA-NeXT-Video 模型进行了扩展消融研究，使用了 10K/15K 样本，展示了有和没有传感器数据的性能。加粗的数字代表启用传感器后的得分。

Annotations	Sensor	Video-LLaVA					LLaVA-NeXT-Video				
ChatGPT	w/o	0.157	0.564	0.539	0.583	0.429	0.108	0.494	0.491	0.546	0.471
	w/	0.255	0.826	0.757	0.785	0.709	0.172	0.696	0.674	0.641	0.649
Gemini	w/o	0.146	0.521	0.551	0.575	0.394	0.108	0.471	0.546	0.546	0.445
	w/	0.299	0.897	0.818	0.833	0.729	0.243	0.896	0.804	0.831	0.717
ChatGPT +Gemini	w/o	0.146	0.512	0.554	0.574	0.430	0.145	0.532	0.534	0.602	0.458
	w/	0.238	0.861	0.770	0.810	0.639	0.208	0.820	0.755	0.806	0.688

Appendix B
视频模型的消融研究

对 Video-LLaVA 和 LLaVA-NeXT-Video 模型进行附加消融研究，结果总结在表 VI 中，包括有和无传感器信息的情况。与结果部分的讨论相似，这些视频模型被训练成在时间任务上表现良好，但是在卫星图像任务上表现不足。然而，与已经在卫星数据上训练的 Qwen 模型不同，为视频模型的提示中添加传感器信息，相较于没有传感器的基础模型表现出更大的改进。