Automatizirano generiranje multimodalnih skupova podataka iz sintetičkih slika korištenjem vizualno-jezičnih modela

Wachtler, Elena

Master's thesis / Diplomski rad

2025

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Electrical Engineering and Computing / Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:168:794623

Rights / Prava: In copyright/Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2025-03-28**

Repository / Repozitorij:

FER Repository - University of Zagreb Faculty of Electrical Engineering and Computing repozitory

zir.nsk.hr

DIGITALNI AKADEMSKI ARHIVI I REPOZITORIJI

UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

MASTER THESIS No. 759

AUTOMATED GENERATION OF MULTIMODAL IMAGE-TEXT DATASETS FROM SYNTHETIC IMAGES USING VISION-LANGUAGE MODELS

Elena Wachtler

Zagreb, February 2025

UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

MASTER THESIS No. 759

AUTOMATED GENERATION OF MULTIMODAL IMAGE-TEXT DATASETS FROM SYNTHETIC IMAGES USING VISION-LANGUAGE MODELS

Elena Wachtler

Zagreb, February 2025

UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Zagreb, 30 September 2024

MASTER THESIS ASSIGNMENT No. 759

Student:	**Elena Wachtler (0036526354)**
Study:	Computing
Profile:	Computer Science
Mentor:	assoc. prof. Tamara Petrović, PhD

Title:	**Automated Generation of Multimodal Image-Text Datasets from Synthetic Images Using Vision-Language Models**

Description:

Vision-language models (VLMs) are trained on datasets containing images and their associated descriptions so that they can perform tasks involving the interpretation and generation of content in both modalities, thus improving the ability to interpret complex scenarios. The master's thesis will focus on developing methods for using vision-language models to generate multimodal pairs (image-text) for training other VLMs. The primary task is to create a process where a pre-trained VLM takes synthetic images generated in 3D modeling tools like Blender or Unreal Engine and generates rich, descriptive textual annotations. The study will evaluate whether these automatically generated datasets can improve the performance of other VLMs. Attention will be given to refining and controlling the description generation process, with a focus on avoiding bias, ensuring diversity, and preventing repetition.

Submission date: 14 February 2025

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 30. rujna 2024.

DIPLOMSKI ZADATAK br. 759

Pristupnica:	**Elena Wachtler (0036526354)**
Studij:	Računarstvo
Profil:	Računarska znanost
Mentorica:	izv. prof. dr. sc. Tamara Petrović

Zadatak: **Automatizirano generiranje multimodalnih skupova podataka iz sintetičkih slika korištenjem vizualno-jezičnih modela**

Opis zadatka:

Vizualno-jezični modeli (VJM) treniraju se na skupovima podataka koji sadrže slike i njihove pridružene opise kako bi mogli obavljati zadatke koji uključuju tumačenje i generiranje sadržaja u oba modaliteta, čime se poboljšava sposobnost tumačenja složenih scenarija. Cilj ovog diplomskog rada je razviti metode za korištenje vizualno-jezičnih modela u svrhu generiranja multimodalnih parova (slika-tekst) za treniranje drugih modela. Glavni zadatak je dizajnirati proces u kojem unaprijed trenirani VJM obrađuje sintetičke slike generirane u alatima za 3D modeliranje, poput Blendera ili Unreal Enginea, i generira bogate i deskriptivne tekstualne anotacije. Rad će istražiti mogu li ovi automatski generirani skupovi podataka poboljšati performanse drugih VJM-ova. Poseban naglasak bit će na usavršavanju i kontroli procesa generiranja opisa, s ciljem izbjegavanja pristranosti, osiguravanja raznolikosti i sprječavanja ponavljanja.

Rok za predaju rada: 14. veljače 2025.

Firstly, I would like to express my gratitude to my mentor, Dr. Sc. Tamara Petrović, for her guidance and support throughout my academic path, from my Bachelor's programme to the present. I would also like to thank Univ. Mag. Ing. El. Techn. Inf. Antonella Barišić Kulaš, who holds the most credit for introducing me to the field of vision-language models. She has been a constant source of motivation and has continued to offer extensive assistance during the development of this thesis. Finally, I am extremely blessed to have so many friends and family to thank for their endless support and patience throughout my academic journey.

Contents

1 Introduction **3**
  1.1 Background and Motivation 3
  1.2 Problem Statement 3
  1.3 Research Questions and Hypotheses 4
  1.4 Contributions 5
  1.5 Structure of the Thesis 6

2 Related Work **7**
  2.1 Vision-Language Models 7
    2.1.1 VLM applications 8
    2.1.2 LLaVA 8
  2.2 Data-centric ML 10
  2.3 Prompt engineering 10

3 Methodology **12**
  3.1 Overview of the Approach 12
    3.1.1 Image Captioning with VQA 12
    3.1.2 Metadata Generation 14
  3.2 Dataset Description 15
    3.2.1 VisDrone Dataset 15
    3.2.2 Dense Dataset 16
  3.3 Ground Truth Data Collection 17
  3.4 Model and Pipeline 18
  3.5 Evaluation Metrics 18

4 Results and Discussion **20**

4.1 Experimental Setup 20
4.1.1 Pipeline 20
4.2 Implementation Details 21
4.3 Image Captioning Results 22
4.3.1 Qualitative analysis 22
4.3.2 Quantitative spatial information analysis 25
4.4 Generated Metadata Results 28
4.5 Discussion and Interpretation of Findings 31

5 Conclusion 34

References 36

Abstract 41

Sažetak 42

A: Examples of Outputs for Image Captioning for Prompts No. 2 to No. 9 . 43

1 Introduction

1.1 Background and Motivation

The word multimodal implies multiple modalities — in particular, in context of artificial intelligence, it mostly refers to data that includes textual and visual information. Most commonly, this involves datasets where images are paired with corresponding text. This approach combines visual information available in images to be further enriched with contextual information provided through text.

Synthetic data are a good source for data that are usually not easily accessible or collected, such as aerial images. However, using synthetic data has some limitations. Some highlight that even though humans may not be able to distinguish synthetically generated from real data, artificial intelligence models, like convolutional neural networks, can, resulting in what is known as the Sim2Real gap [1]. When this gap is translated to aerial imagery, it is a phenomenon referred to as Sim2Air gap. Given the lack of aerial perspective visual data, resorting to synthetic data in applications with unmanned aerial vehicles (UAVs), or drones, is a logical choice. In this work, the capacity of vision-language models to bridge the Sim2Air gap is put to the test.

1.2 Problem Statement

Vision-language models (VLMs) represent a rapidly advancing field in artificial intelligence, combining visual and textual modalities to enable a deeper understanding of images. While these models have shown promising results in tasks such as image captioning, object detection, and visual reasoning, their potential applications in robotics — specifically aerial robotics — remain largely unexplored.

The main mode of information for unmanned aerial vehicles is visual data, which could be significantly enhanced with textual descriptions to improve situational awareness, navigation, and decision-making. This raises several key questions about how VLMs can be effectively integrated into drone applications, including how VLMs can be used in robotics, particularly for aerial systems, what types of captions are necessary to maximize the benefit of VLM-generated metadata, as well as what characteristics make textual descriptions most beneficial in the context of drone imagery.

The answers to these questions are crucial in determining how VLMs can be adapted for real-world applications, such as search-and-rescue missions, environmental monitoring, and surveillance. Furthermore, this research seeks to address whether existing vision-language models can be repurposed for drone-related tasks and how their output can be tailored through prompt engineering to optimize performance.

Cai et al. emphasize that existing datasets are typically small and focus primarily on urban scenes, neglecting rural and industrial areas, which limits the effectiveness of models in processing diverse drone imagery [2]. This is why this work employs synthetic data, aiming to evaluate the performance of VLMs on synthetic aerial imagery.

1.3 Research Questions and Hypotheses

To guide this study, the following research questions were formulated to systematically investigate the effectiveness and potential of automated multimodal dataset generation using vision-language models:

- Can VLMs enhance UAV operations by providing meaningful textual descriptions of aerial imagery?
- To what extent can existing vision-language models, trained on generic datasets, be repurposed for UAV imagery without significant modifications?
- Is it possible to generate high-quality text-image pairs using existing VLMs, thereby reducing the need for manually annotated datasets?
- How does prompt design influence the level of spatial information captured in the responses?

- How can the quality and information content of generated captions be evaluated in a structured manner?
- Can VLMs provide meaningful textual descriptions of synthetic aerial imagery?
- Can the proposed dataset generation pipeline be scaled effectively for large-scale applications?

Based on these research questions, the following hypotheses are proposed:

- UAVs can benefit from vision-language models from their informative captions.
- Existing models, despite being trained on diverse datasets, can be repurposed for UAV imagery.
- New text-image pairs can be created in an automated manner by using existing models, and can serve as an alternative to human-annotated data.
- Carefully designing the prompts can increase spatial information in captions.
- Quality and information value in generated captions can be evaluated using predefined metrics.
- VLMs can successfully provide textual descriptions and identify objects in synthetic images.
- The developed dataset generation pipeline is scalable.

By addressing these research questions and testing these hypotheses, this study aims to assess the suitability of using VLMs for multimodal UAV-related dataset creation, ultimately contributing to the broader understanding of their applicability in aerial robotics. Furthermore, it seeks to evaluate the performance of VLMs when applied to synthetic data.

1.4 Contributions

As VLMs continue to gain popularity, their variety of applications increases rapidly. Some of these applications are further presented in Section 2, and include tasks such as image captioning used in many fields. However, as of the time of this study, little research has been conducted on their use in aerial robotics. While various computer vision models have been widely applied in this domain, VLMs have yet to be thoroughly investigated.

This work introduces a pipeline that, to our knowledge, is the first of its kind to use the vision-language model LLaVA for aerial imagery, with a set selection of prompts. It features prompts with various potential real-world applications, including disaster intervention, drone navigation, data analysis to alleviate human workload, and other.

1.5 Structure of the Thesis

Chapter 2 places this work in the scientific community, first giving an outline of what VLMs are, as well as listing some of their applications in robotics. This chapter also presents data-centric machine learning, as well as prompt engineering. It highlights the relevance of the paper, as well as the novelty it brings. Chapter 3 offers an overview of the approach, describes the datasets, how the data was collected, describes the pipeline and which metrics are used to evaluate the obtained results. In Chapter 4 more details are given about the experimental setup, used hardware, implementation, and the results are presented, evaluated and discussed. Chapter 5 gathers findings, lists limitations of the study, identifying room for improvement, and indicating future work. The Appendix A holds supplementary examples of the model's responses.

2 Related Work

2.1 Vision-Language Models

Xiang Li et al. define vision-language models as a family of artificial intelligence models that combine computer vision and natural language processing techniques to formulate a comprehensive understanding of both visual and textual information [3]. Owing to their integrated architecture combining visual and language processing, unlike their predecessor, vision models, they can not only recognize objects in an image, but also understand the relationships between them. The language model component enables not only solving tasks that output text, such as visual question answering, image captioning, and others, but also allows VLMs to get a better grasp at understanding the context and semantics of visual content.

Feilong Chen et al. [4] divide architectures of vision-language pre-training models into two groups from multimodal fusion perspective: single-stream and dual-stream; and from the overall architectural design perspective, they divide them into encoder-only and encoder-decoder models. Figure 2.1 shows the structure of a typical vision-language model. The models usually consist of an image encoder, with an embedding projector, and text decoder stacked. The embedding projector, which is often a dense neural network, aligns image and text representations [5]. In this architecture, during pre-training, the image encoder and text decoder are frozen, while the multimodal projector is trained. On the other hand, during fine-tuning, both the multimodal projector and text decoder are trained, while the image encoder remains frozen.

Since it was first presented, Jeremy Howard and Sebastian Ruder's ULMFiT (Universal Language Model Fine-tuning) [6] method has been accepted and widely used in many language models. This learning paradigm, including pre-training, fine-tuning and

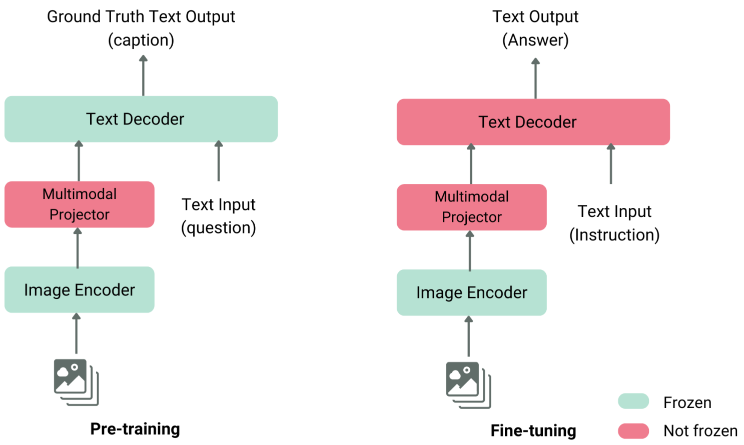


Figure 2.1: Structure of a typical VLM [5]

prediction is employed in VLMs as well, and further extended to combine visual images, delivering the desired results [7]. Hence, for this work, a trained LLaVA model (further explained in 3.1 and 3.4) was used and no additional training was done.

2.1.1 VLM applications

VLMs can be used for a variety of tasks, varying in context, user group, input and output types, and many other characteristics. Some of these tasks are image captioning, visual question answering, visual storytelling, scene understanding, action recognition, visual object reasoning, and many others. Additionally, some VLMs can output images, not only text, which enables them to perform tasks like image segmentation, object detection, and other image editing and transformation tasks. An example of such a VLM is Florence 2 [8].

This wide range of tasks enables VLMs to be further used in diverse real-world applications, from assistive technology, multimedia retrieval, surveillance, to autonomous driving and robotics.

2.1.2 LLaVA

Figure 2.2 depicts a higher level architecture of the LLaVA model. CLIP (**C**ontrastive **L**anguage-**I**mage **P**re-training) [10] is a model developed to learn visual concepts from both images and their corresponding textual descriptions. While traditional image models typically train an image feature extractor alongside a linear classifier to predict labels,

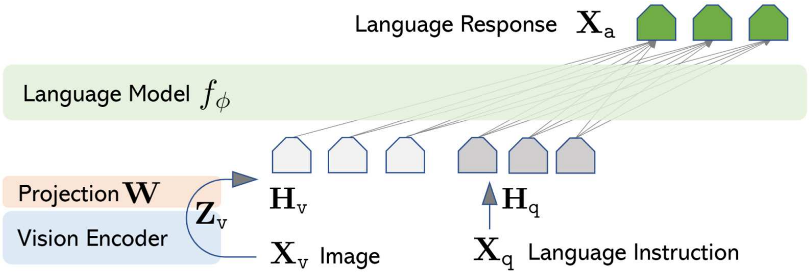


Figure 2.2: LLaVA scheme taken from their official website [9]

CLIP instead trains an image encoder and a text encoder together to predict the correct pairings of image and text. LLaVA's image encoder is based on CLIP's vision encoder.

One of many perks of VLMs' architecture is their modality — having a vision encoder as one, and text decoder as another component, one of them can be replaced without the need for changing the other. LLaVA is a great example of this, where there are versions released with Vicuna [11, 12] or Mistral [13] as a language model, as well as Nous Hermes 2 - Yi-34B [14] for multi-language tasks.

LLaVA can easily be used in pipelines, such as enhancing image-to-image generation by having LLaVA generate textual descriptions of input images that, when fed into the image-to-image generation pipeline, result in a stronger resemblence to the input image [15]. In some work, LLaVA was also connected with SAM2 (Segment Anything 2) [16] to achieve dense grounded understanding of images and videos [17]. In [18], they used depth as additional information in images, testing spatial understanding with vision-language models, in embodied artificial intelligence in particular.

A version of LLaVA, LLaVA-OneVision has been proven effective in chart, diagram and document understanding, it can recognize graphical user interface screenshots of an iPhone and provide operational instructions to search for and open an app, and many other tasks that include image and text. Since this version can process video data as well, it has been tested on a variety of tasks that include video processing, such as providing instructions on how to create a target video based on a static image, analysing differences between two video sequences, analysing and interpreting multi-camera video footage from self-driving cars, and others [19].

Moreover, efforts have been done to further improve LLaVA as a model, with some reporting it being the first single model that can simultaneously push performance boundaries of open large multimodal models in tasks of single-image, multi-image, and video scenarios [19].

2.2 Data-centric ML

Introduction to data-centric AI, a course held in MIT, defines data-centric AI as an emerging science focused on developing techniques to improve datasets. It emphasizes that enhancing data quality is often the best way to improve performance in practical machine learning applications [20]. Singh et al. define data-centric AI as a collection of techniques designed to systematically improve the quality of data used in machine learning pipelines. Unlike model-centric approaches that focus on improving the algorithms or models themselves, data-centric AI recognizes that high-quality data is essential for optimizing the performance of AI systems [21].

In this study, the focus is on creating quality data rather than training or fine-tuning models. The aim was to examine the quality and effectiveness of using vision-language models to describe real-world, as well as synthetic aerial images.

2.3 Prompt engineering

Prompt engineering has gained importance in recent years, with extensive research supporting claims of its effectiveness [22, 23], while others highlight the limitations of this approach [24]. Surprising findings were presented by [25] when prompt engineering is used in vision-language models. They found indications that prompt engineering with LLaVA-NeXT does not alter the attention distribution across modalities, with the context not receiving increased attention after the prompt suggested the model to do so. They investigated how *confusing* the model by giving it either complementary or contradictory context can affect model accuracy and attention. Interestingly, they found that changes in accuracy caused by complementary and contradictory context did not reflect in the attention attribution.

K. Amara et al. also investigated if adding textual description of the image would enhance model performance and confidence, and found indications that models like

LLaVA are already effective at extracting information from the image alone, and that additional textual input may cause confusion due to redundancy or inconsistencies [25]. This indicates that the model's semantic understanding of context originates from the pre-trained component, not the prompt itself. Keeping this in mind was very valuable when designing the prompts.

3 Methodology

3.1 Overview of the Approach

As popularity of vision-language models rises, so does the number of good-quality trained models being published daily. However, many of them are commercial and not available to the broader public for free. This is why LLaVA (**L**arge **L**anguage-**a**nd-**V**ision **A**ssistant) was chosen for this work. It is an open-source, state-of-the-art end-to-end trained large multimodal model that connects a vision encoder and LLM (Large Language Model) for general-purpose visual and language understanding [9].

The proposed pipeline involved selecting a sample from a dataset and processing each image individually. For each image, a corresponding textual prompt, depending on the task, was provided to the model. Finally, the generated responses were qualitatively and quantitatively evaluated.

The main focus of this work was generating two types of responses: captions of images in the form of visual question answering, and image metadata.

3.1.1 Image Captioning with VQA

Visual question answering is a downstream task of correctly answering a question, given a visual input [4]. It is usually in form of a classification task with the model choosing the most suitable answer from multiple options. However, in this case the model is asked for a free-form response, more similar to image captioning, but with a textual prompt included. Chen et al. further state: "Generating relevant and explanatory captions for a visual input requires not only a rich knowledge of language, but also a consistent understanding of scenes, entities, and their interactions appearing in the visual input" [4]. Hence, to test the capability of the model to show understanding of scenes and entities, image captioning was selected as a task.

For the image captioning task, nine prompts were carefully designed and were used in a visual question answering (VQA) setting. Five of them asked for responses that were testing specific applications, while the other four focused on spatial information that can be drawn from the image. The first five prompts were thoughtfully tailored to explore the types of responses they generate and to determine whether they could be useful in real-world applications. Potential use cases include disaster response, assisting drone pilots in navigation, surveillance, as well as some other scenarios where interpreting aerial imagery is essential. Specifically, the first and fourth prompt requested generally descriptive responses, while the fifth prompt also focused on potential challenges in UAV navigation. The second prompt aimed to identify obstacles for UAV flight, while the third prompt was more centered on surveillance tasks. The prompts vary in length, scope, focus and topic, as presented in Table 3.1.

Prompt #	**Prompt Text**
1	This image is an aerial photo taken by a drone. Please provide detailed information about the image contents, focusing on where people are or could be.
2	Can a drone fly undisturbed in the area shown in the image? If not, what are possible problems?
3	What activities are happening in the image? Is there anything dangerous or alarming?
4	Describe the photo that shows an aerial view seen by a drone, highlighting information that could be useful for the drone.
5	Analyze the image from an aerial perspective captured by a drone. Identify any areas of interest, including potential challenges for drone navigation.
6	Describe the spatial arrangement of objects in the image.
7	Describe the layout and spatial relationships of the objects in the image.
8	Describe the image with a focus on the arrangement of objects, such as their relative positions and orientations.
9	You are an expert in mapping and navigation with drones. Describe spatial relations in the given image taken by a drone.

Table 3.1: List of prompts used for image captioning

For each of the images from the sample, a set of five responses was generated. As will be detailed in Section 4.2, a specific parameter ensured that each run would result in a new random response.

3.1.2 Metadata Generation

Public datasets have played a crucial role in advancing deep learning, and their influence is undeniable. Constructing a quality dataset requires time and effort, as collecting, filtering, and annotating are very extensive tasks. To maximize the value of existing data, it can be further enriched with additional information. One such example is metadata creation. Metadata can incorporate diverse elements, including image resolution, camera specifications, but also UAV altitude at the time the image was taken, weather conditions, scene type, human presence, and more. In this work, VLMs are used to accomplish this otherwise time-consuming and labor-intensive task.

For the metadata generation task, a single prompt was designed. This prompt targets information that could be useful for pilots of unmanned aerial vehicles, or drones, asking the model to label each image for the following categories: environment, which could help assess the terrain and potential hazards; weather, which is crucial for planning safe drone operations, as weather conditions directly affect flight safety and performance; and the presence of humans, as identifying human activity is important for ensuring safety, as well as respecting privacy. As was the case with the previous task of image captioning, the model was prompted with each image and the prompt five times, resulting in five responses that were generated. As expected for this kind of prompt and response generation task, some responses were identical despite setting the `do_sample` parameter to `True`. However, this does not pose any issues, as repeated responses were accounted for during processing and interpreted as the model expressing higher confidence in its output. This was the prompt used for generating metadata:

Prompt You are a drone expert. You want to create metadata about the photo you're given. Your response should have the following format:
- environment: desert/forest/city/rural/sea/mountains
- weather: rain/snow/sunny/foggy/clear/overcast/nighttime
- humans: yes/no

Each of the attributes should have one of the labels assigned to it. Make sure to provide a label for each of the attributes.

The above prompt design clearly explains the desired format, as well as the labels for each of the categories. It emphasizes the importance of providing a label for all categories. This should help minimize the number of prompts that will be filtered out due to improper formatting or incomplete information.

3.2 Dataset Description

Drone imagery presents certain unique challenges, including small objects, occlusions, and low resolution. Furthermore, due to the lack of diverse drone training data, accurate training of deep learning detection algorithms has been pointed out as an open challenge [26]. This is why this work focuses on testing the effectiveness of a vision-language model that has demonstrated strong performance with other types of images, putting its generalizability to the test.

For the purpose of this work, two publicly available datasets were used to evaluate the performance of the chosen VLM, LLaVA. These datasets provide diverse visual scenarios captured from an aerial perspective, which makes them suitable for tasks that require an understanding of spatial relationships, object detection, and scene interpretation from above. By having LLaVA generate responses to prompts that include these images, the goal was to analyse its ability to generate meaningful captions and metadata relevant to real-world applications such as drone navigation, surveillance, disaster response, and others.

3.2.1 VisDrone Dataset

VisDrone-DET is a dataset collected for Vision Meets Drone Object Detection in Image Challenge, a challenge held in conjunction with the International Conference on Computer Vision (ICCV) [27]. Starting in 2018, the Vision Meets Drone challenge was held for several years, as part of conferences like ECCV, ICCV, PRCV, as well as CCDM. Specifically, the challenge workshops were organised in conjunction with ECCV 2018 and 2020, ICCV 2019 and 2023, PRCV 2022, and others [28]. Each year, the dataset with the same images was used [29].

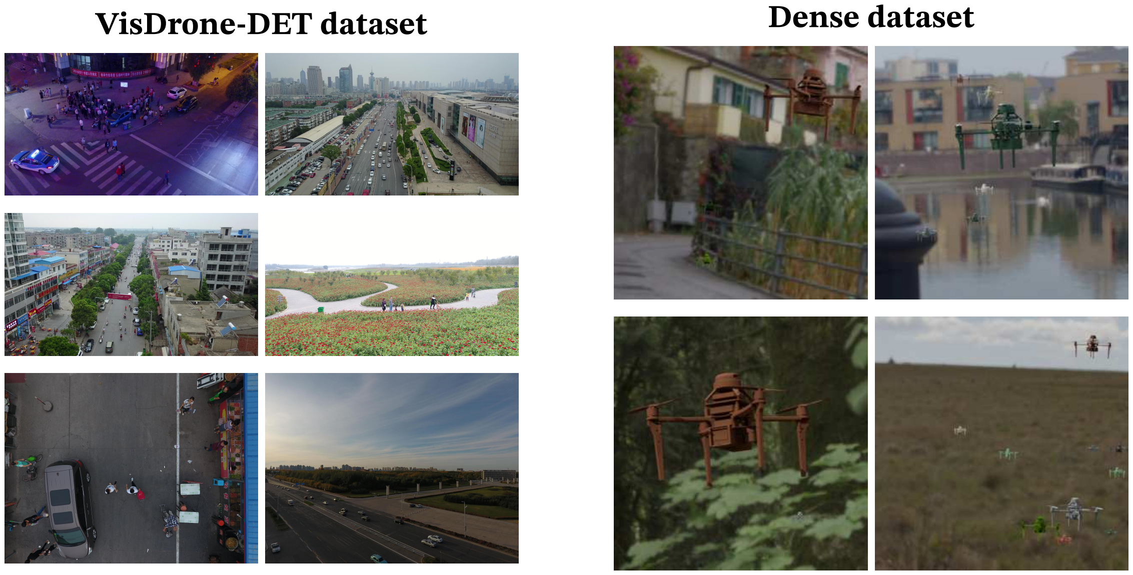
VisDrone benchmark dataset consists of over 280 video clips formed by over 260,000 frames and over 10,000 static images, captured by various drone-mounted cameras. The images vary in location, as they were taken in fourteen different Chinese cities kilometers apart, they capture different objects, including pedestrians, vehicles, bicycles and others, with different density from sparse to crowded scenes (see Figure 3.1). Camera-carrying drones varied in their model, and so did the weather and lighting conditions. Manually obtained annotations target bounding boxes of certain objects are also available, as well as some attributes to provide for better data usage, such as scene visibility, object class and occlusion [30].

In this work we use VisDrone-DET dataset, a subset of the VisDrone dataset that was used for the challenge of object detection in images. This task aims to detect objects of ten predefined categories (pedestrian, person, car, van, bus, truck, motor, bicycle, awningtricycle, and tricycle, where a human maintaining a standing pose or walking is classified as pedestrian, and person otherwise) from individual images taken from drones [27]. Other tasks divide the dataset to VisDrone-VID — a task similar to the one in VisDrone-DET, but the objects are required to be detected in videos, VisDrone-SOT — single-object tracking, a task that aims to estimate the state of a target, indicated in the first frame, in the subsequent video frames, VisDrone-MOT — multi-object tracking, aiming to recover the trajectories of objects in each video frame, and VisDrone-CC — crowd counting, a task to count persons in each video frame [30].

As all VisDrone datasets, the VisDrone-DET is divided into training, validation and testing datasets, with training having 6,471 images, validation 548 and testing 1,580 images. The train dataset is then further divided into annotations and images. For the purpose of this work, only a sample of images in the training dataset was selected. Other than sampling, no preprocessing of this dataset was needed.

3.2.2 Dense Dataset

Dense dataset is a realistic synthetic dataset of 16000 images with large numbers of drones [31]. It was created to test a bio-inspired visual relative localization method for large swarms of drones, using an approach that accentuates shape-based object representation by applying texture randomization [1]. The images were generated using 272 HDRI maps as backgrounds, containing both natural and urban environments with diverse lighting conditions. Each image has a resolution of 300×300 pixels. The im-



VisDrone-DET dataset

Dense dataset

Figure 3.1: Examples of images from the datasets. Left group shows images from the VisDrone-DET dataset, while the group on the right presents examples from the Dense dataset.

ages used for the dataset crops were initially generated with optical parameters matching those of the Intel RealSense D435 color camera, which were mounted on the MRS F450 UAVs[32]. A 3D model of the UAV was created in Blender [33] and used for data generation. Ground-truth labels for the dataset images were generated using the known drone poses. Labels are also available for standard object detection algorithms. From the Dense dataset, only the images were used in this work. Similar to what was done with VisDrone-DET, a sample of images was drawn from the Dense dataset, with no additional preprocessing.

3.3 Ground Truth Data Collection

For comparison with the metadata generated using the LLaVA vision-language model with VisDrone images, the same sample of images was annotated by three individuals from diverse professional backgrounds, serving as ground truth. Specifically, they were given three multiple choice questions for each of the images from the VisDrone-DET sample. These questions asked them to label images in three categories (environment, weather and humans) with the same options as was stated in the prompt the model was provided with. The options were exclusive, meaning each category could only be assigned one label. The label that received the majority of responses from the three annotators was designated as the final answer. In the event of a draw, the first label provided was selected as the final answer.

3.4 Model and Pipeline

The model used here is LLaVa-NeXT, also sometimes referred to as LLaVa-1.6. It has many improvements to the previous LLaVa-1.5, including, but not limited to increasing the input image resolution to at least 672×672, 336×1344, and 1344×336 pixels (which allows it to grasp more visual details and accommodate images of various high resolutions with its AnyRes technique), better visual reasoning and OCR capability, as well as better visual conversation for more scenarios, covering different applications. The latter leads to better world knowledge and logical reasoning [34]. The checkpoint used in this work, `llava-v1.6-mistral-7b-hf`, uses Mistral-7B [35] as its language model specifically, and is available on HuggingFace [13]. Other LLaVA-NeXT checkpoints use different language models, such as Vicuna-1.5 7B and 13B [36], and Nous-Hermes-2-Yi-34B, a state of the art Yi fine-tune [14], which has better commercial licenses, and bilingual support [13]. Compared with other open-source LMMs such as CogVLM or Yi-VL, LLaVA-NeXT achieves the best performance, while in the domain of commercial ones, it even catches up to Gemini Pro, proving it truly is state-of-the-art [34].

3.5 Evaluation Metrics

When choosing evaluation metrics for caption generation through visual question answering, several important notes were taken into account. Traditional measures, such as cosine similarity, would evaluate similarity between the prompts and the responses, taking into account only the textual part of the input, ignoring the image that the prompt is about. However, without data that would include ground truth about answers to the prompts, evaluating response accuracy was not feasible, as it would inevitably include inspecting each of the responses manually.

This is why two measures were employed for image captioning — qualitative analysis, and spatial phrases frequency. To demonstrate the qualitative analysis, several example responses were chosen and analysed. To evaluate spatial information, spatial phrases frequency was defined as a measure of occurrences of phrases used to describe arrangement and spatial relationships between objects. Specifically, it was calculated by the following formula:

$$SF = \frac{N_s}{N_t} \tag{3.1}$$

SF represents the spatial frequency, defined as the ratio of the number of spatial phrases in a response (N_s) to the total number of words in the response (N_t), as shown in Equation (3.1).

Aside from qualitative analysis, generated metadata were evaluated by accuracy, calculated as number of labels in LLaVA-generated metadata that match human annotated labels divided by total sample size 200. The generated responses were filtered in a way that only those responses that provide a valid label from the pre-defined set of labels for each of the categories were taken into account. Additional descriptions and reasoning of label choice were ignored, and only the categories with their labels were extracted for further evaluation.

4 Results and Discussion

4.1 Experimental Setup

All code was written and tested using the Jupyter framework [37]. Code was written and run with Python [38] version 3.11.11 on the first, and 3.12.3 on the second device. There are no differences in the versions that would suggest any disparity in the libraries used or code that was run.

All computationally intensive tasks, including prompting the model, as well as collecting the responses for both captioning and metadata generation, were performed on a high-performance device. The system was equipped with an Intel® Core™ i7-10700 CPU running at 2.90 GHz, featuring eight cores and sixteen threads. It also utilized an NVIDIA GeForce RTX 3090 graphics card with 24 GB of memory, alongside 64 GB of RAM, to efficiently process large-scale computations. The operating system was Ubuntu 20.04.6 LTS. On the other hand, some of the lightweight tasks, such as performing calculations over the previously generated responses, as well as generating graphs, were performed on HP OMEN Laptop 15-en0xxx with an AMD Ryzen™ 7 4800H with Radeon™ Graphics processor with 8 cores and 16.0 GiB of memory, and an NVIDIA GeForce GTX 1660 Ti Mobile (with 6 GB of memory), running on Ubuntu 24.04.1 LTS.

4.1.1 Pipeline

Python library Random [39] was used to randomly draw a sample of 200 out of 6,471 images of the original training dataset of VisDrone-DET. The same library was used to draw a random sample of 200 from 16000 synthetic images of Dense dataset. For both image captioning and metadata generation tasks, the two dataset samples were input together with their prompts. Each prompt-image pair was sent to the model five times, generating five separate responses. Upon generation, the responses were stored to a json file that could easily be processed later.

4.2 Implementation Details

As previously mentioned, LLaVA-1.6 with Mistral-7B is a specific version of the LLaVA model family that was used. This model is available on HuggingFace, which provides code for manipulating the model, including its loading, setting parameters, as well as adapting the prompt to the required prompt structure, and finally prompting the model and generating responses [40].

As each checkpoint is trained with a specific prompt format, depending on the underlying large language model backbone [40], the same prompt structure was followed when prompting. To ensure the prompt is correctly formatted, the `apply_chat_template` method, predefined in `LlavaNextProcessor` of HuggingFace's `transformers`, was applied to the prompts before sending them to the model. In the used checkpoint, the template is as follows: `"[INST] <image>\nWhat is shown in this image? [/INST]"`, where `What is shown in this image?` is the prompt text.

When calling the `model.generate` method, several parameters were set. Specifically, the `max_new_tokens` parameter, that corresponds to the maximum number of tokens to generate, ignoring the number of tokens in the prompt, was set to 500. Secondly, the `do_sample` parameter was set to `True`, meaning sampling was used instead of greedy decoding. This ensures every run results in a new response.

The execution time for metadata generation was 24 minutes for the synthetic Dense dataset, and 26 minutes for the real-world VisDrone dataset. During this time the model processed 200 images, and generated a metadata response five times for each image, which is 1,000 responses in total for each dataset. On average, it took 1.44 seconds to generate a response for an image from Dense, and 1.56 seconds for an image from VisDrone.

The total runtime for generating image captions with VQA for the synthetic dataset was 20 hours and 22 minutes, during which the model processed 200 images, each evaluated with nine prompts and executed five times per prompt-image pair, resulting in a total of 9,000 generated responses. This means that it took around 8 seconds on average to generate a response. For the VisDrone dataset, running one prompt five times for all 200 images took 3 hours and 15 minutes, averaging in around 12 seconds per response. Python library `time` from Python standard library [38] was used to measure execution time.

4.3 Image Captioning Results

4.3.1 Qualitative analysis

Qualitative (manual) analysis showed promising results with responses in captioning, strongly suggesting they could potentially be used in applications the prompts were tailored to test — disaster response, assisting drone pilots in navigation and surveillance, as suggested in 3.1.1. Qualitative analysis is conducted only for Prompt No. 1, while example responses for the remaining prompts are provided in Appendix A.

Figure 4.1 shows one of the five responses the model generated for Prompt No. 1, for a randomly chosen image. The model accurately identifies the presence of humans and vehicles in the image. It successfully detects other object as well, such as buildings, pavement, trees, gates and other. Moreover, the highlighted portions of the output indicate that the model understands the semantic meanings of the identified objects. This emphasizes the advantage of vision-language models over traditional vision models, as the language component provides valuable contextual information that enables VLMs to provide semantic, context-understanding interpretation of given visual input.

Comparable results were observed with images from the synthetic Dense dataset. This is a response to Prompt No. 1 for image shown in Figure 4.2:

> `"The image provided is an aerial photograph taken with a drone. It appears to depict an area with residential properties. We can observe:\n\n1. A multi-story building on the left side of the image with a yellowish facade. The structure has a flat roof and visible shutters on the windows.\n\n2. A wall made of what looks like wooden planks or panels, which could be part of a property’s structure or a separate fence/wall along the side.\n\n3. A green, lush garden area in front of the building,` **`suggesting a well-maintained outdoor space`**`.\n\n4. On the right side of the image, there appears to be a second building or`

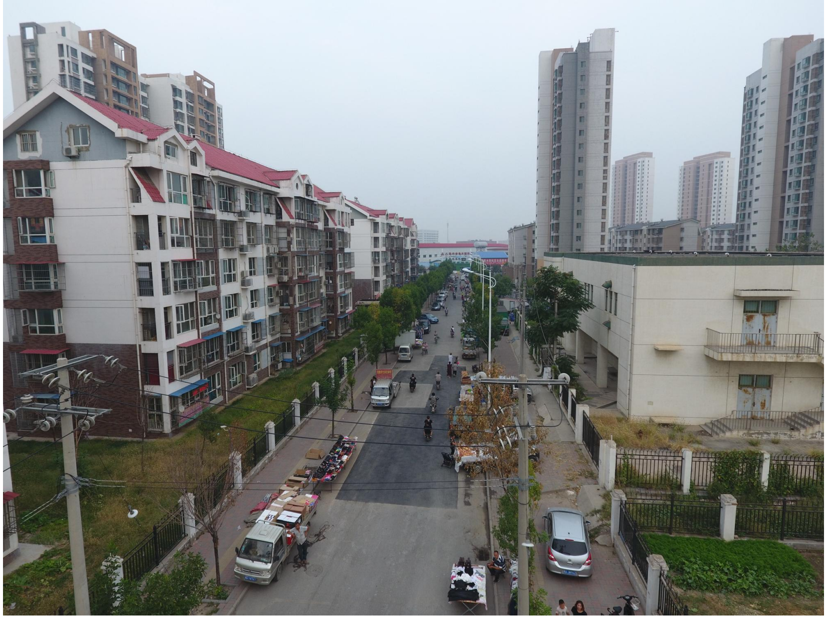


Figure 4.1: An image from VisDrone, randomly selected response to Prompt No. 1: "In the given aerial photo, the view is from a bird's-eye perspective, looking down a long street lined with residential buildings on both sides. The buildings have a series of balconies and rooftops, **and their architecture suggests a modern design, possibly in an Asian country**. The roofs are partially decorated with terracotta elements in shades of red and brown.\n\nOn the right side of the street, there is a clear view through to an urban area with high-rise buildings, **indicating a commercial or business district**. The sky is overcast, **suggesting a cool or possibly cloudy day**.\n\nIn the middle of the image, there is an open paved space **which appears to be a traffic divide** or an unoccupied parking lot with some vehicles parked on it. Across the street from this area, numerous individuals are present, some standing alone, while others are gathered in small groups. **Their presence could indicate a nearby public space, like a park or a community area, accessible from the street**.\n\nNear the end of the street, there is a section of road with no traffic visible at the time of the photo. On each side of the street, a row of small vegetation or trees borders the buildings, providing a touch of green to the urban landscape. On the right side of the street, a low fence follows the boundary and is punctuated with gates **that likely lead to the individual residential compounds or private property**.\n\nOverall, the image captures a typical day in a suburban area with a mix of residential and commercial zones, and the presence of people **suggests a lively neighborhood**."

```
structure visible, although only its outline is clear due to the distance
and angle.\n\nGiven the limited information available, no definitive
conclusions can be drawn about the presence or activity of people in
this scene. The aerial perspective provides a bird's-eye view of the
area but does not provide direct evidence of human activity or presence."
```

Even though the image is significantly lower in resolution with only 300 × 300 pixels, the model still recognizes the main objects. It showed good understanding of the image context, and expressed uncertainty in determining human activity.

For the same image, shown in Figure 4.2, and prompt, another response indicated se-

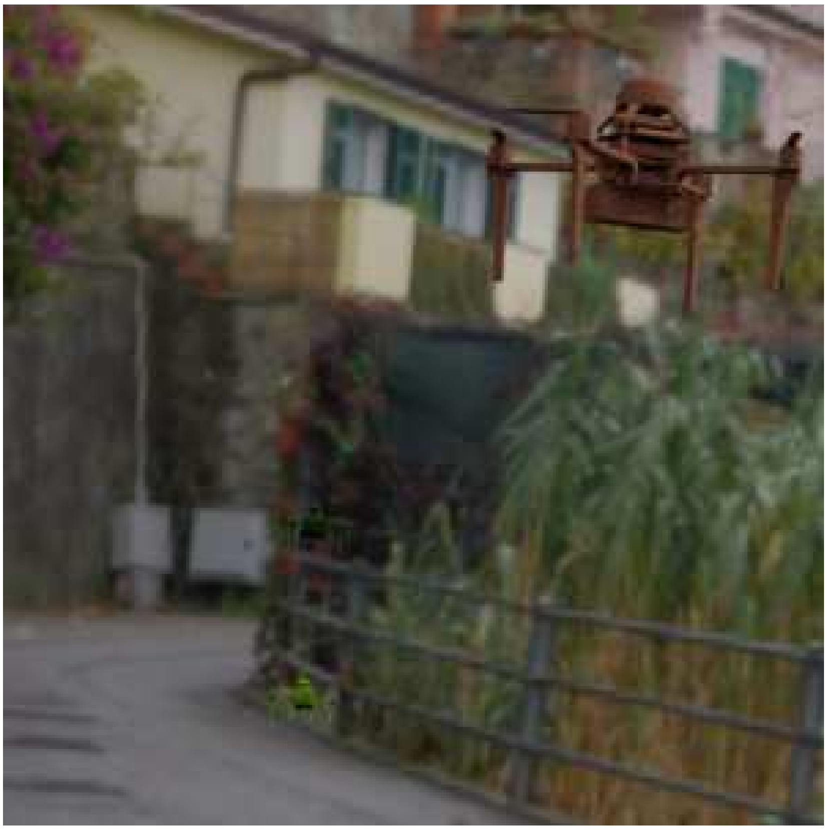


Figure 4.2: An image from Dense

mantic understanding of the environment of the image, correctly assessing the presence of people not by identifying them directly, but by observing the image context:

> `"(...) The overall setting seems to be a suburban neighborhood with a mix of residential properties and vegetation. There are no people visible in the image, so the question of where people could be is not directly answerable from this photo.\n\nHowever,` **`considering the residential nature of the area, one would expect to see people in one or more of the surrounding houses`**`. If this were a real-life scenario, and not a drone photograph of an abandoned building as you have described, it is` **`reasonable to assume that people might be inside the homes or possibly attending to tasks in the front gardens or yards`**`. (...)"`

However, the model output is flawed, as it does say "if this were a real-life scenario, and not a drone photograph of an abandoned building as you have described", when no such information were mentioned in the prompt, which may be considered a type of hallucination.

The model's effectiveness in identifying the drone in synthetic images varied. In this particular example response to Prompt No. 1 for the same image 4.2, it inaccurately described one of the drones as a "large object that resembles a rusted vehicle or machinery part", but successfully identified another drone as "a small, green, humanoid figure that is likely the drone itself":

> "The image features an aerial viewpoint, likely taken by a drone. In the foreground, we see a fence that indicates the boundary of the property. Behind the fence, there is a collection of plants and shrubs, including what appears to be a patch of tall grass and a tree, possibly indicating a landscaped or natural area.\n\nIn the background, slightly elevated, there is a two-story building with a flat roof, visible windows, and some external features such as a drainpipe. The building is partially obscured, suggesting additional structures or landscaping elements out of view.\n\nOn the roof of the building, there is a **large object that resembles a rusted vehicle or machinery part**, which is prominent in the composition. This object spans a significant portion of the roof, indicating it might be a **piece of equipment or a stationary machine**.\n\nIn the bottom right corner of the photo, there is a **small, green, humanoid figure that is likely the drone itself**, providing context for the perspective and scale of the image. This figure is slightly out of focus due to the angle and distance from the drone's camera to it.\n\nThere are no clear signs of people in this photo based on the visible elements. The drone seems to be flying quite high above the area, and there is no indication of human activity on the ground. The image conveys a sense of quietness and solitude, as there are no moving vehicles or individuals visible in this aerial snapshot of an outdoor environment."

In this example, the model incorrectly places the first drone on the roof of the building and the second drone to the right bottom of the image, but performs well in describing the surrounding environment and its function.

Based on empirical observations, the generated captions mostly stay on topic, with almost no instances of meaningless output.

4.3.2 Quantitative spatial information analysis

For the purpose of measuring spatial information present in responses, a set of spatial phrases was defined. This set included the following phrases: above, below, next to, beside, in front of, behind, on top of, underneath, near, far from, adjacent to, between, inside, outside, around, over, across, within, opposite, towards, away from, along, by, through, amid, surrounded by, alongside, close to, beneath, overhead, under, in the vicinity of, bordering, at the edge of, at the center of, in the middle of, atop, on the side of, across from, perpendicular to, parallel to, nearby, cornered by, in proximity to, aligned with, level with, diagonally opposite, offset from, anchored to, at the base of, right, left, bottom, top.

Figures 4.3 and 4.4 show the average count and frequency of spatial phrases in each of five generated answers to each of nine prompts, across the 200 images that were sampled from VisDrone dataset. Similarly, figures 4.6 and 4.7 show the count and frequency of spatial phrases in each of five generated responses of each of nine prompts, when given the sample of 200 images from the synthetic Dense dataset. Both count and frequency are shown because some responses are significantly longer than others, which makes the spatial phrases count insufficient for accurate analysis.

Firstly, in figures 4.3 and 4.6 it is easy to notice that Prompt No. 3, which asks the model to focus on activities happening in the image and identify any potentially dangerous or alarming elements, invokes responses with a lower count of spatial phrases. Given that this was a prompt tailored to mimic one that would be used in emergency response situations, and that the model focuses on actions and activities, this kind of behaviour is understandable. Moreover, neither the VisDrone nor Dense dataset are expected to contain data with alarming or dangerous scenery that was targeted in the prompt. This results in generally shorter responses, which can be noted from the relationship of spatial phrases count with the frequency.

Secondly, in Figure 4.4 it can be observed that aside from the spatial prompts (Prompts No. 6 to No. 9), Prompt No. 1 also has high spatial frequency. This can be explained with the fact that, even though it was designed to test applicability in certain scenarios rather than gather spatial information, the prompt asks for the position of humans in the image, which inherently involves spatial data.

Figure 4.5 shows the average spatial frequency of the first five prompts (blue) contrasted to the average spatial frequency of the last four prompts (orange), where the x-axis represents the individual images. It highlights the difference in the spatial frequency

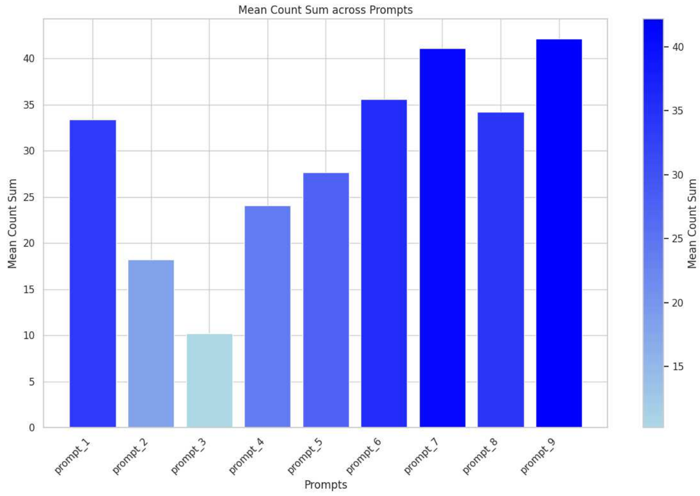


Figure 4.3: Count of spatial phrases in responses for each prompt for VisDrone sample

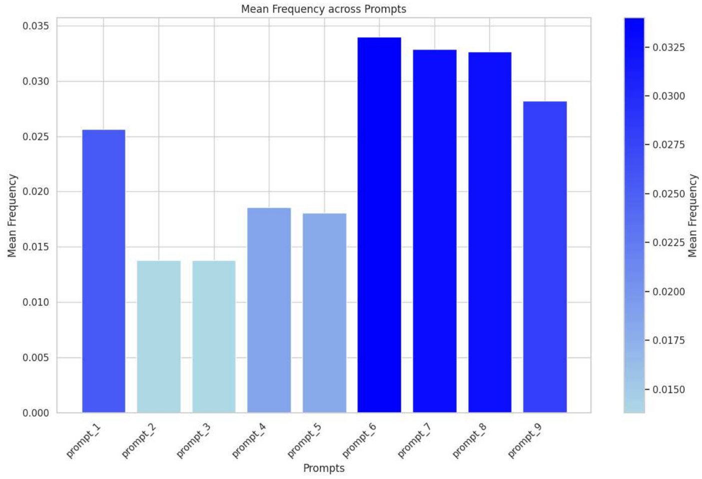


Figure 4.4: Frequency of spatial phrases in responses for each prompt for VisDrone sample

that can be observed in prompts that targeted spatial information, as opposed to those that were more general or had different application. This proves that a carefully designed prompt can enhance spatial information in responses. Figure 4.8 depicts that using synthetic images shows the same trend.

Similarly to Figure 4.5, in 4.8, it is clearly visible that the spatial phrases frequency is significantly higher in prompts that targeted spatial information, once again proving that prompts that ask for spatial information do get more spatial information in the response.

The same trend of spatial prompts drawing more spatial information in responses was observed in real-world images (VisDrone dataset) and synthetic images (Dense dataset).

Although there are slight differences in spatial phrase frequency between the spa-

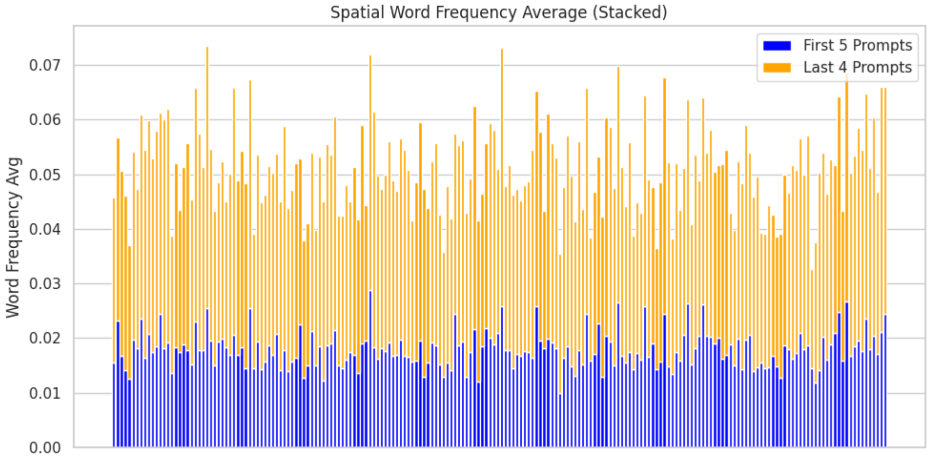


Figure 4.5: Comparison of spatial phrase count and frequency across prompts for VisDrone sample

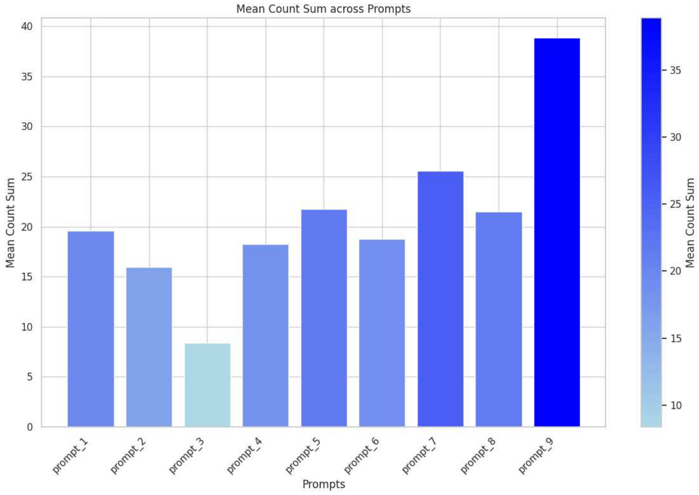


Figure 4.6: Count of spatial phrases in responses for each prompt for the synthetic sample

tial prompts, these differences are inconsistent between the two datasets, and are not substantial enough to be considered significant.

4.4 Generated Metadata Results

Consistent with the methodology employed in processing human annotated metadata, from the five responses generated by LLaVA, the label that was in most responses was taken as the final answer. Ambiguous answers that contained multiple labels for a category, or offered a label that was not provided, were ignored. Draws were handled with the first provided selected as final. Environment category was proved most successful, with an accuracy of 93%. Humans category had an accuracy of 69.5%, while weather was lowest with 62%, as shown in table 4.1. Environment category resulted in great ac-

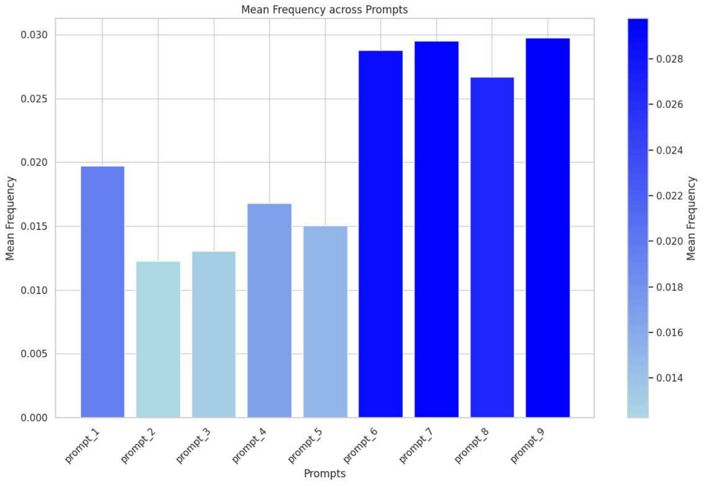


Figure 4.7: Frequency of spatial phrases in responses for each prompt for the synthetic sample

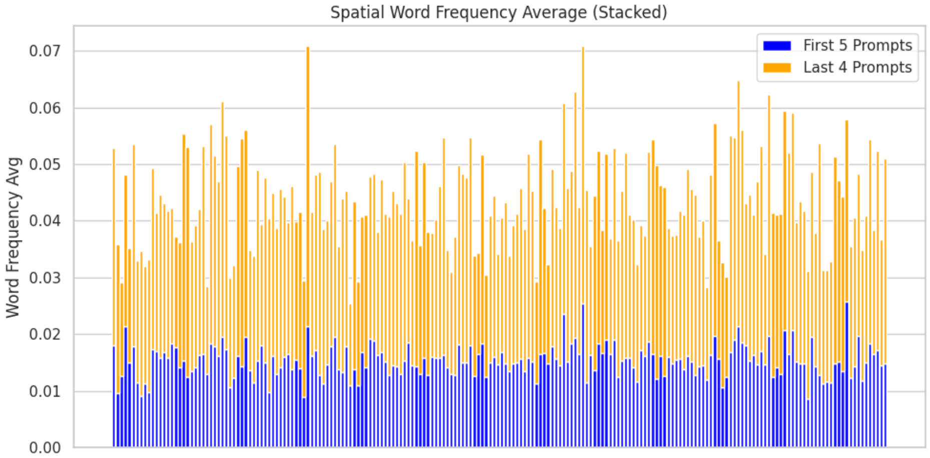


Figure 4.8: Comparison of spatial phrase count and frequency across prompts for the synthetic sample

curacy, while the lower accuracy of humans category can be explained by ambiguity in determining humans' presence in the scene. However, the result for weather was not expected to be so low.

After the labeling process, some annotators reported difficulty distinguishing between labels `clear` and `sunny`. Further analysis of the model-generated metadata showed that the model often had mixed answers for the same labels. When these labels were merged into a single label, the accuracy for weather category increased to 87.5%. Initially, these two labels were employed to offer information that could be relevant to applications including drones, as sunny weather can introduce challenges with direct sunlight degrading the drone's camera vision. While the reasoning for selecting these particular labels was justified, the notably lower accuracy score highlights the significance of care-

Category	**Accuracy (%)**
Environment	93
Humans	69.5
Weather (original)	62
Weather (after merging labels)	87.5

Table 4.1: Accuracy of different categories before and after merging two labels for the weather category

fully selecting label categories, as such decisions directly impact evaluation in a later stage.

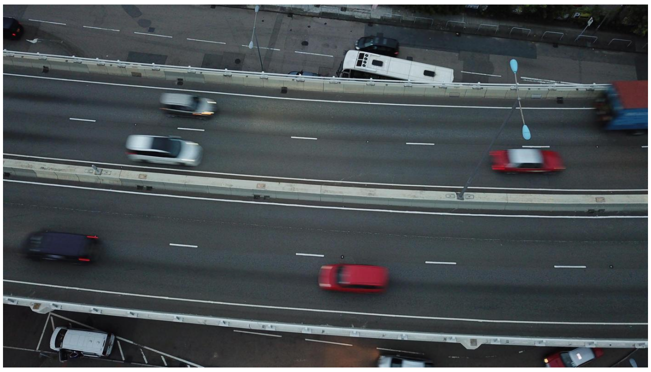


Figure 4.9:

1. Based on the image, here is the metadata:
    - environment: city/urban
    - weather: clear/daytime
    - humans: yes
2. - environment: urban/city/intercity
    - weather: clear/sunny
    - humans: yes (vehicles driving on the freeway)
3. environment: city
    weather: clear
    humans: yes (there are vehicles visible)
4. environment: city
    weather: clear
    humans: yes
5. - environment: city
    - weather: clear
    - humans: no

Figure 4.9 shows an example when annotators and the model disagreed on the humans label. In four out of five responses, the model says there are humans present. Furthermore, in two responses, it offers an explanation for this choice — there have to be humans driving the moving cars. All human annotators labeled this image as no human presence.

Figure 4.10 shows a similar example of metadata where the model differed from ground truth, but offered a reasonable explanation for its choice.

In Figure 4.11 is an image for which the model produced inconsistent classifications regarding human presence, with `no` recorded in three responses and `yes` in two, while the annotators agreed on it being `yes` (there is a person standing next to a bicycle close to the left edge of the image). The model's difficulty in identifying humans in this scene

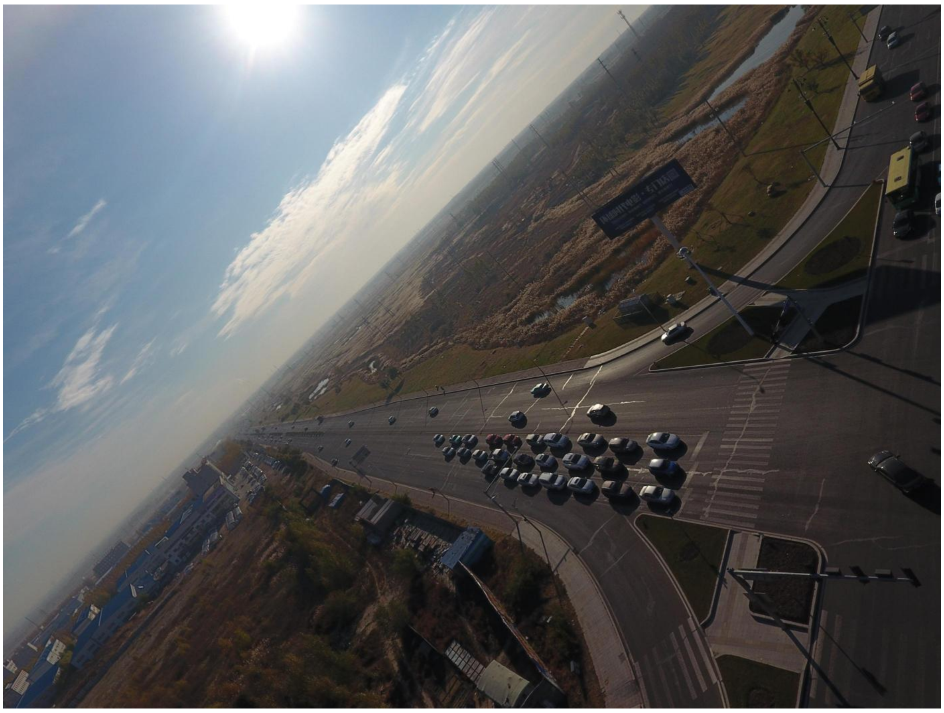


Figure 4.10: "environment: rural\nweather: sunny\nhumans: yes (indicated by the presence of cars on the road, which are typically driven by people)"

may lie in lower illumination due to the nighttime setting, or the high visual complexity of the crowded composition.

Another crowded scene can be seen in Figure 4.12, for which both the annotators and the model lacked consensus, with the model identifying humans in two out of three responses (two responses were discarded due to incorrect formatting or ambiguity in the provided answer), and the annotators in one out of three.

The given examples help clarify why the accuracy is lower in the humans category than it is for the other two categories.

4.5 Discussion and Interpretation of Findings

Empirical evidence has shown the model exhibits deep contextual understanding of functions that the detected objects have in their environment, expanding object detection to semantic understanding of what the objects represent. Even though misidentification does happen, as well as adding off-topic or irrelevant information to the response, the responses mostly provide accurate, relevant information. The model's success in recognizing drones in synthetic images varied, yet it presented capability of describing and rea-

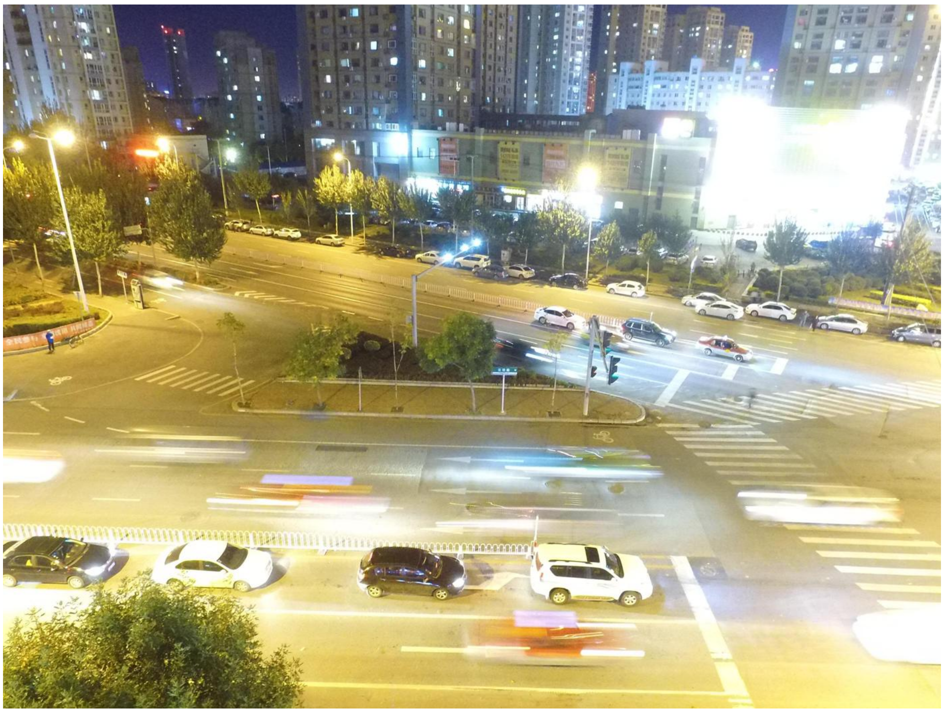


Figure 4.11: Example image with indecisive category labels in model's responses

soning about the surrounding area. Measured spatial phrases frequency proved prompt engineering can increase spatial information in responses, while qualitative analysis of selected examples showed accuracy in describing relationships between objects. The model achieved high accuracy in generating metadata compared to human-annotated data, with certain categories being less successful due to ambiguity, lower illumination, or high visual complexity. Moreover, a trend was observed of the model being indecisive between similar labels as human annotators in certain images, demonstrating its humanlike behaviour.

Some responses present the phenomenon of adding textual description of the image to the prompt *confusing* the model [25], as was the case in one of the outputs showed for the selected image from Dense dataset shown in Figure 4.2.

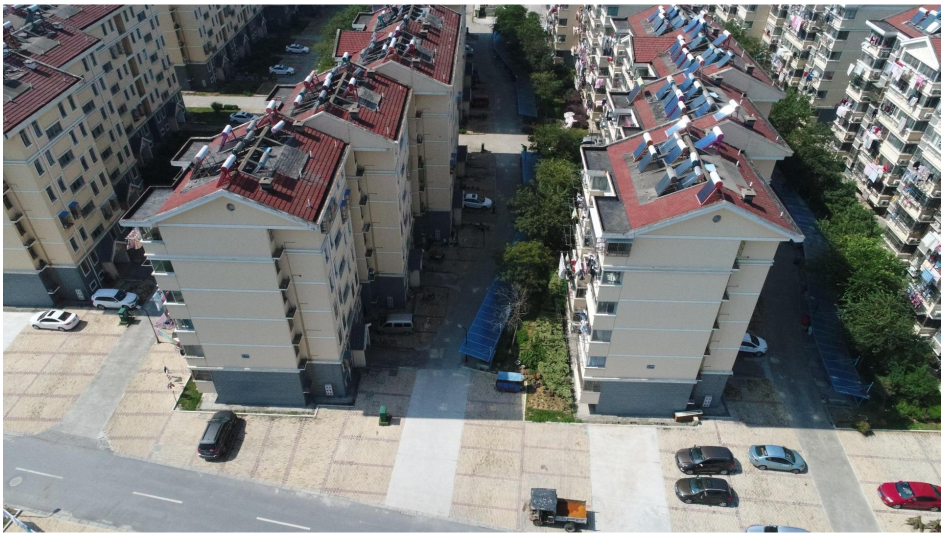


Figure 4.12: Example image with indecisive category labels in both the model's and annotators' responses

5 Conclusion

This work has shown that a selected VLM, LLaVA-1.6 with Mistral-7B language model can be used to describe aerial imagery, despite the fact that it was not specifically designed for it. LLaVA successfully generated meaningful responses to prompts intended to test diverse applications, including UAV pilot assistance through relevant descriptions, aiding emergency response, and surveillance, all based on carefully designed prompts. The model exhibited an understanding of the meaning of objects and their surrounding, demonstrating its reasoning capabilities. By using spatial phrases frequency as an evaluation measure, it was proved that targeting spatial information in prompts can result in enhanced spatial information in the generated responses. However, the proposed spatial evaluation measure did not assess the accuracy in the provided responses. Future work will require a dataset of aerial images with ground truth spatial arrangements of objects to address this limitation.

Furthermore, it was demonstrated that LLaVA could provide relevant descriptions of synthetic aerial imagery. Despite the images being lower in resolution than in the real-world dataset used here, the model still presented capability to effectively recognize objects, and reason about the environment. However, its capacity to identify drones varied, indicating the need for further evaluation using a different synthetic dataset to assess performance across varied conditions. One potential synthetic dataset that could be used to test VLM performance on is **D**epth **E**stimation o**N** **S**ynthetic **E**vents (DENSE), which provides synthetic events and accurate ground truth [41].

The findings of metadata generation task indicate that VLMs can achieve high accuracy in labeling images in predefined categories, proving it could be used as an alternative to manual image annotation. Future research should draw a sample that is not necessarily random, but is carefully selected to adequately represent the full range of possible labels within each category. Additionally, it should incorporate datasets that are more diverse in environment, part of the world they were recorded in, and include ground truth information. Metadata could be expanded to include more categories, depending on the application.

In conclusion, this study provides valuable insights into the performance of a vision-language model on aerial imagery. While the model demonstrated promise in recognizing objects and reasoning about the environment, its performance on the synthetic dataset varied. Future work should include prompts that test for a wider range of applications, as well as more diverse synthetic datasets and enhancing evaluation metrics.

References

[1] A. Barisic, F. Petric, and S. Bogdan, “Sim2air - synthetic aerial dataset for uav monitoring,” *IEEE Robotics and Automation Letters*, vol. 7, no. 2, p. 3757–3764, Apr. 2022. https://doi.org/10.1109/lra.2022.3147337

[2] W. Cai, K. Jin, J. Hou, C. Guo, L. Wu, and W. Yang, “Vdd: Varied drone dataset for semantic segmentation,” 2024. [Online]. Available: https://arxiv.org/abs/2305.13608

[3] X. Li, C. Wen, Y. Hu, Z. Yuan, and X. X. Zhu, “Vision-language models in remote sensing: Current progress and future trends,” 2024. [Online]. Available: https://arxiv.org/abs/2305.05726

[4] F.-L. Chen, D.-Z. Zhang, M.-L. Han, X.-Y. Chen, J. Shi, S. Xu, and B. Xu, “Vlp: A survey on vision-language pre-training,” *Machine Intelligence Research*, vol. 20, no. 1, p. 38–56, Jan. 2023. https://doi.org/10.1007/s11633-022-1369-5

[5] M. Noyan and E. Beeching, “Vision language models explained,” Apr. 2024, accessed: 2025. [Online]. Available: https://huggingface.co/blog/vlms

[6] J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,” 2018. [Online]. Available: https://arxiv.org/abs/1801.06146

[7] Hugging Face, “Introduction to vision language models,” 2025, accessed: 2025. [Online]. Available: https://huggingface.co/learn/computer-vision-course/unit4/multimodal-models/vlm-intro

[8] Microsoft, “Florence-2-large,” https://huggingface.co/microsoft/Florence-2-large, 2023, accessed: 2025-02.

[9] L. Team, “Llava: Large language and vision assistant,” 2024, accessed: 2025. [Online]. Available: https://llava-vl.github.io/

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[11] Liuhaotian, “Llava-v1.6-vicuna-13b,” https://huggingface.co/liuhaotian/llava-v1.6-vicuna-13b, 2023, accessed: 2025-02.

[12] ——, “Llava-v1.6-vicuna-7b,” https://huggingface.co/liuhaotian/llava-v1.6-vicuna-7b, 2023, accessed: 2025-02.

[13] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, “Llava v1.6 mistral-7b,” 2023, lLaVA v1.6 integrates a pre-trained large language model with a vision encoder for multimodal chatbot applications, enhancing upon previous versions with improved data diversity and dynamic high-resolution capabilities. [Online]. Available: https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf

[14] N. Research, “Nous-hermes-2-yi-34b,” 2024, fine-tuned version of the Yi model, trained on 1,000,000 entries of primarily GPT-4 generated data. [Online]. Available: https://huggingface.co/NousResearch/Nous-Hermes-2-Yi-34B

[15] Z. Ding, P. Li, Q. Yang, and S. Li, “Enhance image-to-image generation with llava-generated prompts,” in *2024 5th International Conference on Information Science, Parallel and Distributed Systems (ISPDS)*. IEEE, May 2024, p. 77–81. https://doi.org/10.1109/ispds62779.2024.10667513

[16] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, “Sam 2: Segment anything in images and videos,” 2024. [Online]. Available: https://arxiv.org/abs/2408.00714

[17] H. Yuan, X. Li, T. Zhang, Z. Huang, S. Xu, S. Ji, Y. Tong, L. Qi, J. Feng, and M.-H. Yang, “Sa2va: Marrying sam2 with llava for dense grounded understanding of

images and videos," 2025. [Online]. Available: https://arxiv.org/abs/2501.04001

[18] W. Cai, I. Ponomarenko, J. Yuan, X. Li, W. Yang, H. Dong, and B. Zhao, "Spatialbot: Precise spatial understanding with vision language models," 2024. [Online]. Available: https://arxiv.org/abs/2406.13642

[19] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, and C. Li, "Llava-onevision: Easy visual task transfer," 2024. [Online]. Available: https://arxiv.org/abs/2408.03326

[20] MIT CSAIL Data-Centric AI, "MIT CSAIL Data-Centric AI Initiative," 2025, accessed: 2025-02. [Online]. Available: https://dcai.csail.mit.edu/

[21] P. Singh, "Systematic review of data-centric approaches in artificial intelligence and machine learning," *Data Science and Management*, vol. 6, no. 3, pp. 144–157, 2023. https://doi.org/https://doi.org/10.1016/j.dsm.2023.06.001

[22] B. Chen, Z. Zhang, N. Langrené, and S. Zhu, "Unleashing the potential of prompt engineering in large language models: a comprehensive review," 2024. [Online]. Available: https://arxiv.org/abs/2310.14735

[23] H. Strobelt, A. Webson, V. Sanh, B. Hoover, J. Beyer, H. Pfister, and A. M. Rush, "Interactive and visual prompt engineering for ad-hoc task adaptation with large language models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 1, pp. 1146–1156, 2023. https://doi.org/10.1109/TVCG.2022.3209479

[24] A. Webson and E. Pavlick, "Do prompt-based models really understand the meaning of their prompts?" 2022. [Online]. Available: https://arxiv.org/abs/2109.01247

[25] K. Amara, L. Klein, C. Lüth, P. Jäger, H. Strobelt, and M. El-Assady, "Why context matters in vqa and reasoning: Semantic interventions for vlm input modalities," 2024. [Online]. Available: https://arxiv.org/abs/2410.01690

[26] C. Li, S. C. Sun, Z. Wei, A. Tsourdos, and W. Guo, "Scarce data driven deep learning of drones via generalized data distribution space," 2022. [Online]. Available: https://arxiv.org/abs/2108.08244

[27] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. Zhang, L. Bo, H. Shi, R. Zhu, A. Kumar, A. Li, A. Zinollayev, A. Askergaliyev, A. Schumann, B. Mao, B. Lee, C. Liu, C. Chen, C. Pan, C. Huo, D. Yu, D. Cong, D. Zeng, D. Reddy Pailla, D. Li, D. Wang, D. Cho, D. Zhang, F. Bai, G. Jose, G. Gao, G. Liu, H. Xiong, H. Qi, H. Wang, H. Qiu, H. Li, H. Lu, I. Kim, J. Kim, J. Shen, J. Lee, J. Ge, J. Xu, J. Zhou, J. Meier, J. Won Choi, J. Hu, J. Zhang, J. Huang, K. Huang, K. Wang, L. Sommer, L. Jin, and L. Zhang, “Visdrone-det2019: The vision meets drone object detection in image challenge results,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[28] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, “Detection and tracking meet drones challenge,” 2021. [Online]. Available: https://arxiv.org/abs/2001.06303

[29] Y. Cao, Z. He, L. Wang, W. Wang, Y. Yuan, D. Zhang, J. Zhang, P. Zhu, L. Van Gool, J. Han, S. Hoi, Q. Hu, and M. Liu, “Visdrone-det2021: The vision meets drone object detection challenge results,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021, pp. 2847–2854.

[30] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, “Detection and tracking meet drones challenge,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380–7399, 2021.

[31] M. Křížek, M. Vrba, A. B. Kulaš, S. Bogdan, and M. Saska, “Bio-inspired visual relative localization for large swarms of uavs,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2024, p. 11825–11831. https://doi.org/10.1109/icra57147.2024.10610100

[32] D. Hert, T. Baca, P. Petracek, V. Kratky, V. Spurny, M. Petrlik, M. Vrba, D. Zaitlik, P. Stoudek, V. Walter, P. Stepan, J. Horyna, V. Pritzl, G. Silano, D. Bonilla Licea, P. Stibinger, R. Penicka, T. Nascimento, and M. Saska, “Mrs modular uav hardware platforms for supporting research in real-world outdoor and indoor environments,” in *2022 International Conference on Unmanned Aircraft Systems (ICUAS)*. IEEE, Jun. 2022, p. 1264–1273. https://doi.org/10.1109/icuas54217.2022.9836083

[33] B. O. Community, *Blender - a 3D modelling and rendering package*, Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. [Online]. Available: http://www.blender.org

[34] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," January 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-01-30-llava-next/

[35] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023. [Online]. Available: https://arxiv.org/abs/2310.06825

[36] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," 2023. [Online]. Available: https://arxiv.org/abs/2306.05685

[37] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, and C. Willing, "Jupyter notebooks – a publishing format for reproducible computational workflows," in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, F. Loizides and B. Schmidt, Eds. IOS Press, 2016, pp. 87 – 90.

[38] Python Software Foundation, "Python standard library, version 3.x," 2024. [Online]. Available: https://docs.python.org/3/library/

[39] G. Van Rossum, *The Python Library Reference, release 3.8.2.* Python Software Foundation, 2020.

[40] H. Face, "Llava-next model documentation," 2023, accessed: 2025-02. [Online]. Available: https://huggingface.co/docs/transformers/v4.49.0/en/model_doc/llava_next

[41] J. Hidalgo-Carrió, D. Gehrig, and D. Scaramuzza, "Learning monocular dense depth from events," 2020. [Online]. Available: https://arxiv.org/abs/2010.08350

Abstract

Automated Generation of Multimodal Image-Text Datasets from Synthetic Images Using Vision-Language Models

Elena Wachtler

Vision-language models (VLMs) combine the functionalities of computer vision and language models, which enables them to further enrich visual information available in images with contextual information provided through text. However, general-purpose models are not specifically optimized for aerial imagery, a domain where specialized models are often required. Many resort to using synthetic data in such cases, but this approach has its limitations. In this work, a system using a general-purpose VLM was developed and then used to generate responses based on carefully designed prompts paired with both real-world and synthetic aerial images. The system was evaluated on two tasks: image captioning via visual question answering and metadata generation. Results indicate that the system effectively generates multimodal data for aerial imagery, demonstrating a deeper understanding of context of the visual content, and promising reasoning capabilities. These findings highlight the potential of VLMs in real-world applications such as surveillance, search-and-rescue missions, autonomous navigation, and robotics.

Keywords: vision-language models; VLM; synthetic data; natural language processing; generating multimodal datasets; aerial imagery; UAVs

Sažetak

Automatizirano generiranje multimodalnih skupova podataka iz sintetičkih slika korištenjem vizualno-jezičnih modela

Elena Wachtler

Vizualno-jezični modeli (VJM) objedinjuju funkcionalnosti računalnog vida i jezičnih modela, obogaćujući vizualne informacije iz slika kontekstualnim informacijama koje pruža tekst. Međutim, modeli opće namjene nisu optimirani za fotografije snimljene iz zraka, zbog čega su za tu svrhu obično potrebni specijalizirani modeli. U takvim se slučajevima često koriste sintetički podatci, no takav pristup ima ograničenja. U ovome je radu razvijen sustav koji koristi VJM opće namjene kako bi proizveo odgovore na pažljivo osmišljene upite uparene s kako slikama iz zraka iz stvarnog svijeta, tako i sa sintetičkim slikama iz zraka. Sustav je evaluiran na dvama zadatcima: opisivanje slike odgovorom na pitanje i stvaranje metapodataka. Rezultati ukazuju na učinkovitost sustava u generiranju multimodalnih podataka za slike snimljene iz zraka, pokazujući dublje razumijevanje konteksta vizualnog sadržaja, kao i obećavajuće sposobnosti zaključivanja. Ova opažanja naglašavaju potencijal VJM-ova za primjene u stvarnom svijetu, poput nadzora, akcija traganja i spašavanja, kao i autonomne navigacije te robotike.

Ključne riječi: vizualno-jezični modeli; sintetički podatci; obrada prirodnog jezika, generiranje multimodalnih skupova podataka; slike iz zraka; bespilotne letjelice

Appendix A: Examples of Outputs for Image Captioning for Prompts No. 2 to No. 9

The following two quotes are responses to Prompt No. 2:

```
Can a drone fly undisturbed in the area shown in the image? If not,
what are possible problems?
```

and image shown in Figure 4.1:

```
"In the image you provided, there appears to be a crowded residential
area with densely packed buildings, a busy street with various types of
vehicles, and tall buildings that suggest a developed urban environment.
\n\nFlying a drone in this area may pose several challenges and risks:
\n\n1. **Regulations:** Drone regulations can vary by country and
by local laws. You will need to check if flying a drone is allowed
in the specific area where you plan to fly. Some areas may have
restrictions on drone operation due to safety reasons, proximity to
certain establishments, or other regional requirements.\n\n
2. **Obstacles:** The area is filled with buildings, trees, and other
objects which could be obstacles to drone flight. Buildings can also
cause the signal for the drone’s remote controller to be blocked.\n\n3.
**Pedestrians and Vehicles:** The streets seem to be busy with people
walking and vehicles moving. This can lead to potential collisions or
distract the pilot from managing the drone’s safety.\n\n4. **Privacy
Concerns:** Drones can capture images and record videos, which raises
questions about privacy as these images might inadvertently include
people who do not consent to being photographed. Additionally, flying a
drone without the proper permissions can lead to privacy issues.\n\n5.
**Sound and Vibration:** Drones can produce noise and vibrations,
which might disturb the residents or the occupants of other nearby
buildings.\n\n6. **Interference with Air Traffic Control (ATC):** In
some countries, flying a drone can interfere with ATC, especially in busy
```

urban areas where there might be airspace restrictions.\n\n7. **Safety Concerns:** The presence of a drone in an unregulated area where people or vehicles are nearby can pose safety risks. A drone might be injured by a car or a person running unexpectedly, and there's always a risk of accidental injuries.\n\nIf you decide to fly a drone in this area, it's crucial to assess the risks, abide by local regulations, and take into account the safety and comfort of the people around you. Always prioritize safety and respect privacy when operating drones."

"In the image you provided, the area appears to be a dense urban setting. There are several considerations that could affect whether a drone could fly undisturbed:\n\n1. **Airspace Regulations**: The Chinese government, where the image is likely taken, has strict regulations about drone flight. In urban areas, it is typically required to have special permissions and to follow specific routes and altitudes. Violating these regulations may result in penalties.\n\n2. **Air Traffic Control**: The area shown in the image may be close to an aviation area, where there would be air traffic control regulations requiring a certain distance from airports (in some cases, this distance can be several kilometers).\n\n3. **Infrastructure**: Large buildings block line-of-sight communication between the drone and the pilot, which is crucial to maintaining control. In addition, flying close to buildings or structures could pose a risk if the drone were to collide with them.\n\n4. **Local Laws and Regulations**: There may be local laws regarding where drones are allowed to fly, especially if the flight might cause noise or safety hazards to people or structures on the ground.\n\n5. **Crowards and Public Spaces**: The image shows a street with several pedestrians and vehicles. Flying a drone in a crowded area can pose safety risks and disrupt the daily activities of bystanders.\n\n6. **Stability and Wind Conditions**: Urban areas often have air pollution and turbulence from buildings, which can affect the stability of a drone. Steering a drone in such conditions would require expert maneuvering.\n\n7. **Privacy Concerns**: Drones can breach personal privacy if used for surveillance or photography in public spaces without permission.\n\n8. **Local Wildlife**: Urban environments attract wildlife, which could pose risks if a drone comes too close to nesting sites or animals in flight paths.\n\nGiven these considerations, it would be challenging to fly a drone unintruded in this area without proper permissions and compliance with all relevant local and national regulations."

Another example of output to Prompt No. 2 is given below, for image A.1:

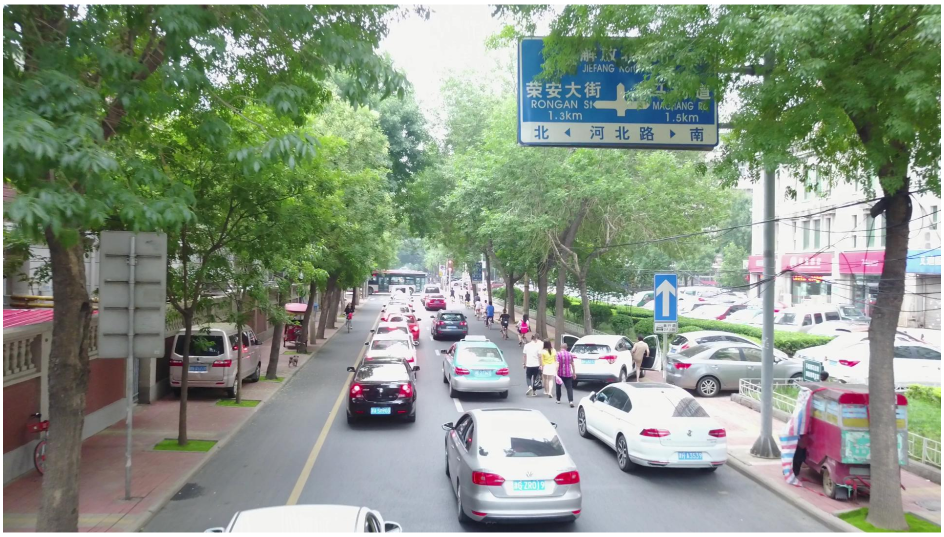


Figure A.1: Example image with potentially interesting content for obstacles

"The area shown in the image appears to be a busy urban street with multiple vehicles, both moving and parked, as well as numerous pedestrians. This creates a challenging environment for drone flight, as there are several obstacles and potential risks that could interfere with the drone's operation:\n\n1. Tall trees and buildings: These can be a challenge because they could cause the drone to get stuck or entangled.\n\n2. Vehicle and pedestrian traffic: The drone operator must be cautious to avoid interfering with vehicle or pedestrian pathways or causing accidents.\n\n3. Obstacles on buildings and street furniture: There are signs, street lamps, traffic lights, and other structures that drones must navigate around.\n\n4. Noise and airborne debris: The noise and airborne debris from vehicles and machinery could impair the drone's sensors and control systems.\n\n5. Regulatory restrictions: Flying drones in busy or crowded areas might be subject to local regulations, and operators should ensure they are in compliance with any laws or ordinances that prohibit or restrict drone flight.\n\n6. Privacy concerns: Flying a drone in a public area with many people around could raise privacy concerns.\n\n7. Signal interference: There could be a risk of signal interference with the drone's controllers if there is sensitive equipment nearby or if the drone is controlled via remote link.\n\nFor these reasons, it would be advisable to plan the drone flight in an area with fewer obstructions and where flight is permitted to minimize risks and disruptions."

Shown below is an answer to Figure A.1 paired with Prompt No. 4:

Describe the photo that shows an aerial view seen by a drone, highlighting information that could be useful for the drone.

```
"The photo depicts a bustling street scene from an aerial perspective.
Here are some observations and pieces of information that could be useful
for a drone operating in this area:\n\n1. **Vehicle Movement:** The road
is full of various types of vehicles including cars, motorcycles, and
buses. Some vehicles are moving in the same direction on the main
road, while others are parked or stopped in what appears to be a bus
or taxi stand.\n\n2. **Pedestrians:** Several people are visible on
the sidewalks and crosswalks. This suggests the presence of pedestrian
traffic that should be navigated carefully by the drone.\n\n
3. **Signage:** The most prominent sign in the view is a large,
blue rectangular sign with Chinese characters, which could indicate
a specific area, landmark, or directional guidance. Other signs are
visible, but they are too small or in too much shadow to fully read
or understand the context.\n\n4. **Trees and Canopies:** The street is
lined with trees and buildings with canopies or awnings providing shade
and protection from the elements. This provides a sense of the urban
environment’s greenery and overall aesthetic.\n\n5. **Infrastructure:**
The presence of power lines is visible, indicating an urban setting with
developed infrastructure.\n\n6. **Density of Buildings and Vegetation:**
The buildings are closely packed, suggesting a densely populated area.
The street appears to have a lot of surrounding greenery, suggesting
green spaces or parks in the vicinity.\n\n7. **Bicycles and Scooters:**
There are some parked bicycles and scooters, indicating that this might
be a city where multi-modal transportation is common.\n\nFor the drone to
safely navigate through such an environment, it would need to be aware of
vehicle movements, pedestrian traffic, and potential obstructions like
trees or signposts. The drone would also need guidance on legal flying
altitudes and adhering to local regulations regarding drone operation
in urban areas. Additionally, the drone could be used for tasks such as
surveying, mapping, real estate photography, or traffic monitoring."
```

Prompt No. 3 aimed at testing whether the model can be used to help surveillance:

```
What activities are happening in the image? Is there anything dangerous
or alarming?
```

Below are several responses for the image showed in A.2, which features elements that could be interesting to point out in a surveillance scenario:

```
"The image shows a busy urban scene at night. There is a significant
crowd of people gathered on a street, with many standing on the corner
at a crosswalk. Various activities appear to be taking place, such as
```

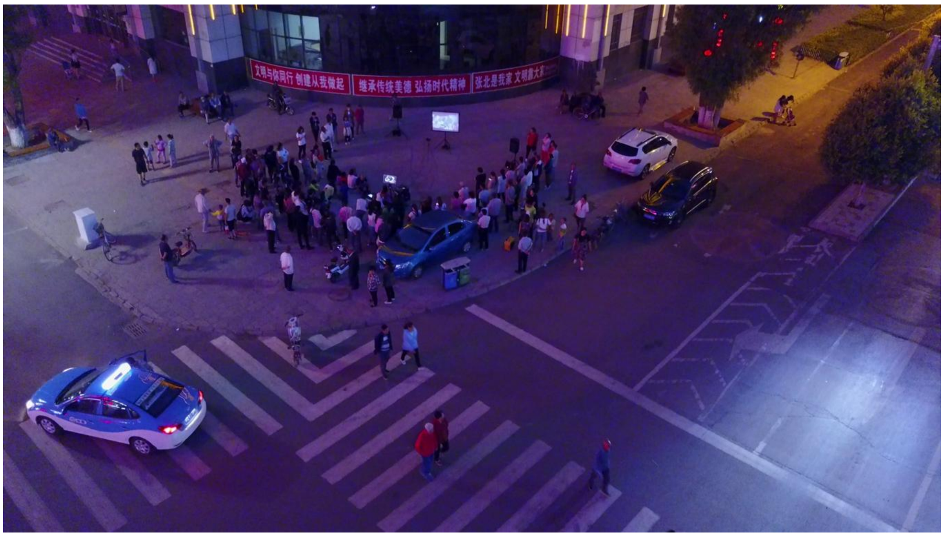


Figure A.2: Example image with potentially interesting content for surveillance

people waiting to cross the street, others socializing or congregating, and potentially some form of event or performance, suggested by the presence of a large screen and a gathering that seems to be around it.\n\nThere are parked cars on either side of the street and a couple of vehicles with their hazard lights on. This could indicate that an incident has occurred or that there is ongoing traffic that is causing caution to be exercised. There is also a police car with its lights activated, which further suggests some sort of emergency or incident is being handled.\n\nFrom a safety standpoint, one alarming element might be the large gathering of people waiting to cross the street. Pedestrians should always be cautious when crossing the road, and it’s important for drivers to be aware of this and for pedestrians to follow traffic signals. Additionally, the presence of emergency vehicles and their active warning signals indicate that there is an ongoing incident that requires attention and may pose hazards to individuals in the area."

"In the image, there is a gathering of people on a city street at night, and the street is busy with several cars parked and possibly moving as well. It appears to be a public event or demonstration, as there’s a large screen displaying what seems to be video content, and the crowd is facing towards it. The people are standing in various orientations, with some facing the screen and others towards the vehicles.\n\nThere is nothing inherently dangerous or alarming visible in the image; the scene appears to be a typical urban gathering where people are engaged in social activities. Additionally, street lighting and the presence of vehicles suggest that the area is securely illuminated, which is standard for city streets to ensure visibility and safety."

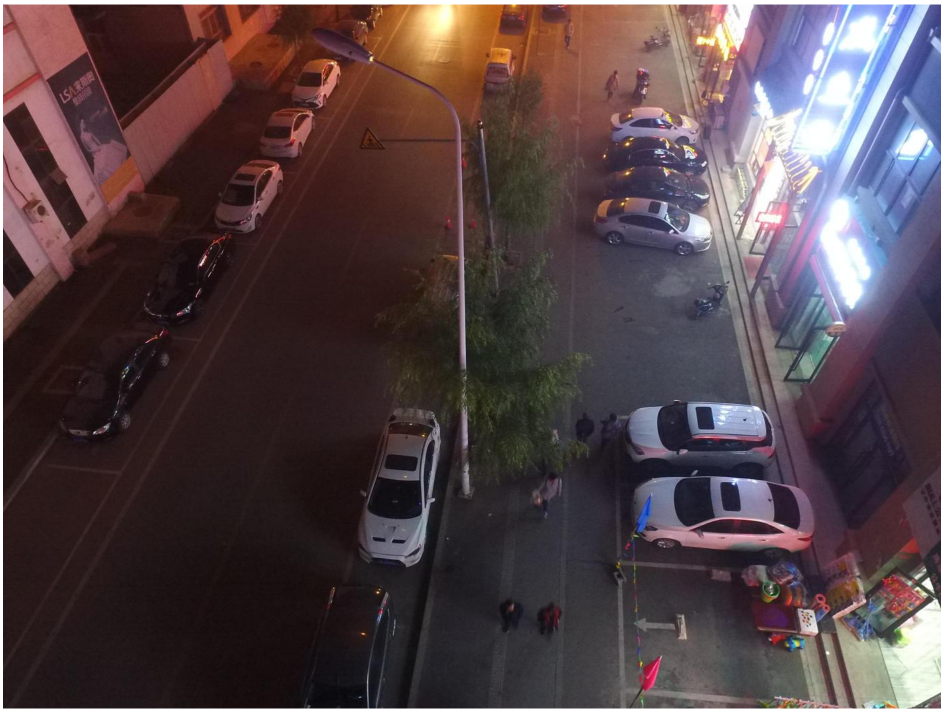


Figure A.3: Example image

An answer to Prompt No. 5:

```
Analyze the image from an aerial perspective captured by a drone.
Identify any areas of interest, including potential challenges for
drone navigation.
```

when paired with image A.3 states:

```
"The image you've provided is taken from an aerial perspective and
shows a street during what appears to be nighttime. Here are some
points of interest and potential challenges for drone navigation within
this scene:\n\n1. The street is lined with parked cars, including a
mix of stationary cars and what looks to be scooters. This could be a
challenge for drone navigation, especially if the drone is operating
at low altitude, as it needs to avoid these obstacles.\n\n2. There are
several pedestrians who are difficult to identify from this height;
drones would need to be cautious to avoid collisions with people.\n\n3.
In the center of the image, there is a tree which could be a visual
reference point for navigational purposes.\n\n4. The lighting in the
area is primarily artificial, which can be a good guide for drone
navigation but could also cause the drone to be drawn towards bright
light sources.\n\n5. The architecture of the buildings suggests a
commercial area, possibly with shops, which can imply challenges like
```

```
strong electromagnetic interference (EMI) that drones are sometimes exposed to.\n\n6. There are a few bicycles, which are also potential obstacles for a drone navigating the area.\n\n7. A motorcycle is parked in the scene, which might be more challenging to avoid than cars due to its smaller size.\n\n8. There is a yellow-lit sign on the right side of the image, but it is not clearly legible from this perspective.\n\n9. There is a line of parked vehicles which might stretch into the distance, making visual navigation more complex.\n\n10. There is a red and white object in the far right of the image, but it is too small to determine its nature.\n\nTo navigate safely, the drone operator would need to be aware of these points of interest and account for them based on the specific capabilities and limitations of the drone being used."
```

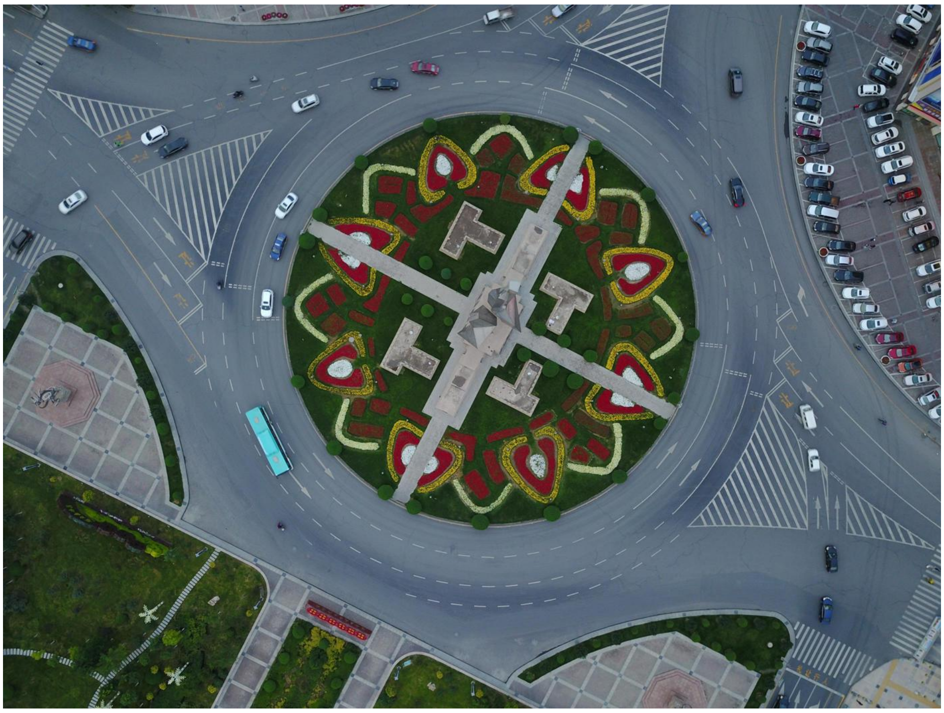


Figure A.4: Example image

Below are a few example responses for Prompt No. 6:

```
Describe the spatial arrangement of objects in the image.
```

paired with image shown in Figure A.4:

```
"The image showcases an aerial view of a roundabout, which is a circular circular area of road for vehicular traffic to circulate around, in an urban setting. In the center of the roundabout, there is a circular flower bed featuring a well-manicured lawn. The lawn is adorned with various large, colorful flowers and plants arranged in a symmetrical
```

> pattern, forming the shape of a heart in the middle.\n\nSurrounding the roundabout is a concentric set of roads that form a series of connected circular arteries. These roads appear to be designed for vehicles to move around the central flower bed, ensuring a smooth flow of traffic.\n\nAt the bottom of the image, there is an inset that provides a bit of a 'before' view of the roundabout, showing a patch of green grass prior to the construction of the roundabout structure.\n\nIn the streets surrounding the roundabout, there are multiple parallel cars parked on both sides of the road, suggesting that the area is a public space designed for parked vehicles. These cars vary in color and model, indicating a diverse range of personal transportation in the area.\n\nThe streets are numbered with zebra crossings at regular intervals, and there are pedestrian walkways connected to the roads at various points. The overall layout of the roads and pedestrian pathways suggests a well-planned urban environment designed to facilitate both vehicular and pedestrian traffic.\n\nThe image does not provide sufficient information to determine whether the specific location of the roundabout or the style of the landscape design is unique or well-known, or if it represents an innovative approach to urban green spaces."

> "The image depicts an aerial view of an intersection with a flowerbed at its center. The intersection is circular in nature and is designed with multiple road elements such as roundabouts, path segments forming a circular loop, lane markings, and paved areas. The flowerbed is planted with various flower varieties and also includes geometric shapes and designs, contributing to its aesthetic appeal.\n\nOn the periphery of the intersection, there are numerous elements including streets, sidewalks, and parking areas. The streets are lined with parked cars, traffic flow is visible, and pedestrians can be seen on the sidewalks. In the top right corner of the image, there is a shopping plaza with parked cars and a clear demarcation of parking bays.\n\nThe overall layout of the intersection is carefully planned to provide smooth traffic flow with the central circular pathway acting as a means of pedestrian and vehicle access from multiple directions."

Another example of a response to Prompt No. 6 is listed below, but paired with an image that does not have such a clear overview of the area, as shown in Figure A.5

> "The image shows an outdoor area with various objects arranged in a somewhat chaotic manner, reflecting a possibly informal or makeshift market or gathering space. Here's a description of the spatial arrangement:

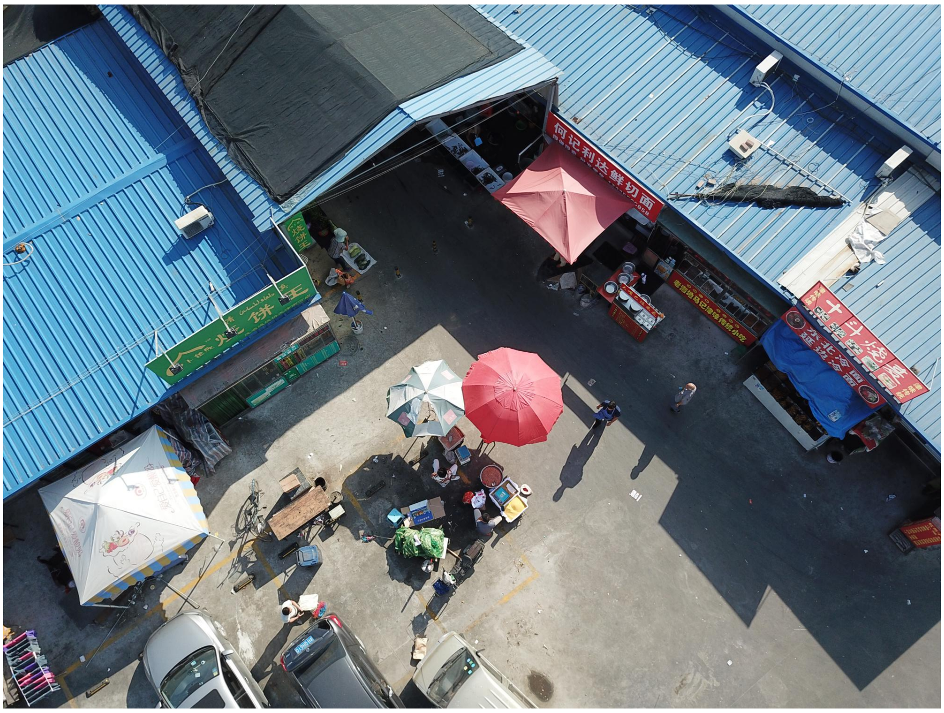


Figure A.5: Example image

```
\n\n1. From the aerial perspective, the center of the image is dominated
by a few colorful umbrellas (or parasols), which are probably used
to provide shade. They are not all open, with one prominently open
in the center, creating a focal point.\n\n2. Under these umbrellas,
there are vendor stalls selling goods. The stalls have signs with
writing that is not English, suggesting a local market or commercial
area.\n\n3. The buildings flanking this area have signs of a similar
nature, indicating the presence of businesses, possibly food stalls,
shops, or services.\n\n4. People are scattered throughout the space,
likely customers, vendors, or onlookers.\n\n5. Parked vehicles are
visible around the perimeter of the area. Cars are on the left side,
and a motorcycle is parked under one of the umbrellas.\n\n6. Trash
is scattered around, hinting at activity and possibly indicating where
food or other disposable items are being consumed.\n\n7. There's a
makeshift stand or kiosk in the center area with a red umbrella, which
seems to overlook the scene, possibly serving a supervisory or oversight
role.\n\nThe layout suggests a community or informal market gathering
space, with stalls haphazardly arranged and a semi-formal structure
provided by the umbrellas. The presence of people and vehicles suggests
activity and usage of the space. It appears to be a sunny day as
evidenced by the vibrant colors and shadows cast."
```

Below are a couple of responses for the same image shown in Figure A.5 and Prompt No. 7:

Describe the layout and spatial relationships of the objects in the image.

"The image shows an open-air marketplace with a collection of covered stands that appear to be small businesses or vendors. These stands are situated on a paved area, possibly a public square or a dedicated market zone, surrounded by what seems to be a commercial block with buildings painted in shades of light to medium blue.\n\nThe market area is bustling with various goods and products on display, with items such as umbrellas, tables, chairs, and other items suggesting a variety of businesses, possibly including food stalls, clothing vendors, or crafts.\n\nDominating the center of the marketplace is a large, red open umbrella, which provides a covered space for people to stand and engage with the vendors. This umbrella stands out due to its size and bright color, serving as a central meeting point or focal point in the market.\n\nIn the background under a second, smaller awning, there are more umbrellas and tables, indicating additional vendors in the area. The presence of multiple umbrellas and tables demonstrates the commercial nature of the space.\n\nTo the bottom of the image, several cars are parked, suggesting that the market area is accessible by vehicle and located in a region where personal transportation is common. The cars are positioned along the curve of the street, parallel to the marketplace.\n\nOverall, the layout of the objects in the image conveys an organized yet vibrant marketplace setup, with items arranged in a way that encourages interaction and engagement among the market's visitors and vendors."

"In the image, we see a busy street scene with a focus on an open-air market area under a corrugated metal building. The market area is sheltered under two large umbrellas, one red and one white, providing shade for potential customers and goods.\n\nSurrounding the market, there are several stalls or small stands set up at the curb, likely for vendors to sell their wares. Overhead, a few more umbrellas are open to provide extra shade or to designate specific stalls.\n\nTo the right of the image, there is a row of buildings, each with its own corrugated metal awnings. Cars are parked along the street, and the ground is populated with various items, including a bicycle and a few trash bins. There is also a person, possibly a customer, standing in the middle of the scene, suggesting some level of activity.\n\nThe layout of the objects within the image suggests a bustling environment with a focus on the market area, where people might come to shop or browse. The parked cars indicate that the location is accessible by vehicle, and the presence of the larger umbrellas suggests that this market may be

> open-air, where vendors can set up their goods and customers can move around more freely."

Another example response to Prompt No. 6, paired with image shown in Figure A.6, is listed below:

> "The image shows an aerial view of an urban area during the day. It consists of several streets lined with commercial buildings on the side. Each street is divided into lanes with parking spots and pedestrian sidewalks. Vehicles are parked along the streets, filling in the available parking spaces.\n\nTrees are planted along the sidewalks and at intervals near the buildings. Power lines and street lights are visible above the streets, and traffic markings are drawn down the center of the roads. The layout suggests a well-organized city plan with dedicated urban spaces for different activities. The buildings have signs with various colors and lettering displayed on their facades, indicating the presence of commercial establishments. The overall scene is characterized by orderly planning and a mix of functionality and aesthetics."

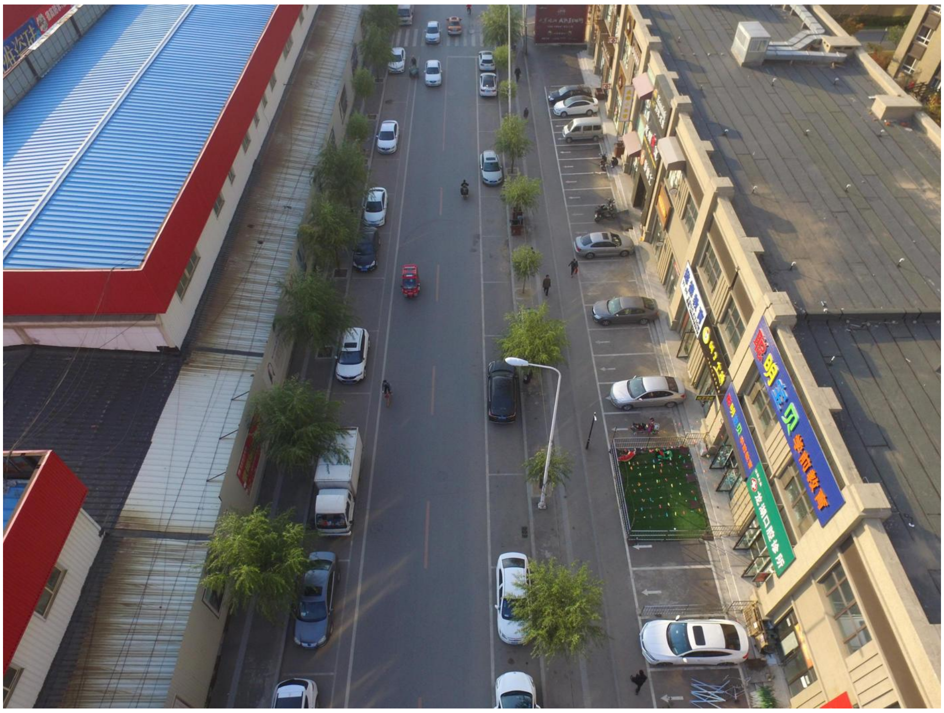


Figure A.6: Example image

Another response was generated for the same image shown in Figure A.6, but for Prompt No. 7:

"The image shows an aerial view of a street scene. On each side of the street, there are buildings with commercial establishments, as indicated by signs, some of which are bilingual. The roads are marked with dashed white lines to indicate lanes. Vehicles are parked along the side of the street, and there are trees planted at intervals, providing some greenery in the urban setting. The street appears to be a one-way road, with cars parked parallel to the curb. On the left side of the road, there are industrial buildings with what appears to be a blue metal roof, while on the right side, there are more traditional commercial buildings, likely shops and restaurants. There's a clear contrast between the industrial area and the more residential and business-oriented right side of the road. The overall layout suggests a well-organized urban environment with a mix of different functions in the adjacent buildings."

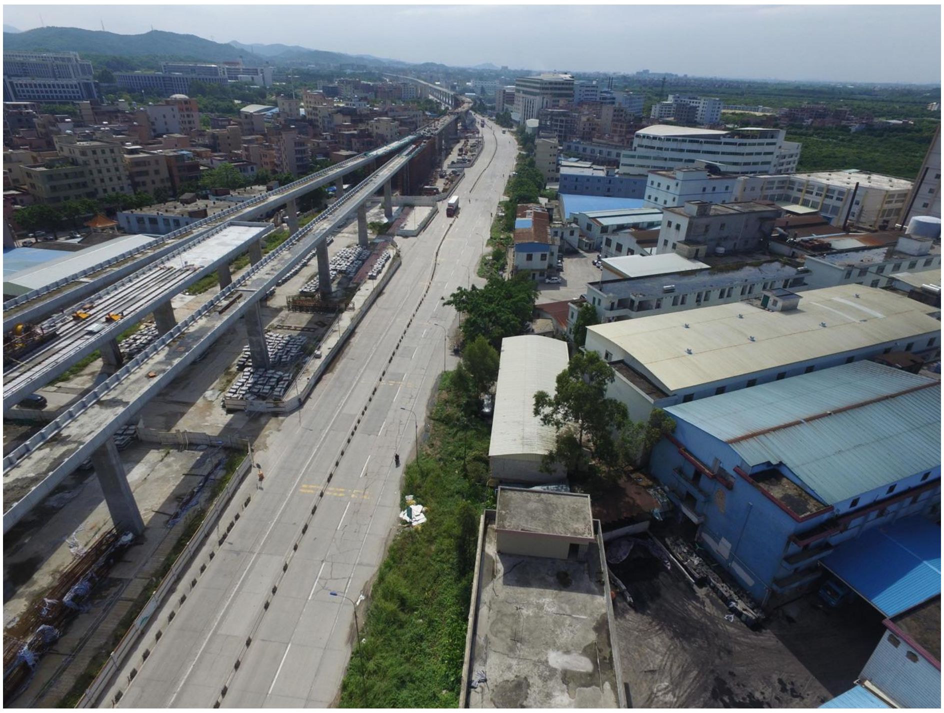


Figure A.7: Example image

The image depicted in Figure A.7, when sent to the model paired with Prompt No. 6, resulted in one of the responses being:

"The image is an aerial view of an urban area with a focus on transportation infrastructure. Here's a description of the spatial arrangement of objects:\n\n1. Expressway (top left side): The expressway runs parallel to the cityscape, with overpasses connecting the different parts of the road.\n\n2. Bridge (top left side): A large bridge spans the width of the urban area, crossing over an empty land area below and connecting with another road.\n\n3. Skyway (top left side): A skyway runs elevated above the expressway, providing additional transportation options.\n\n4. Buildings (throughout the image): The city contains numerous buildings of varying sizes and densities. Some are concentrated along the expressway, while others are interspersed throughout the area.\n\n5. Greenery (throughout the image): Small patches of green land are scattered between buildings and infrastructure.\n\n6. Roads and intersections (bottom center): Below the skyway and bridge, multiple roads intersect, weaving their way around the urban area.\n\n7. Rooftops (throughout the image): The rooftops of the buildings indicate the built-up nature of the city.\n\nThis image captures a modern city’s layout, with a clear delineation between transportation infrastructure and built-up urban areas."

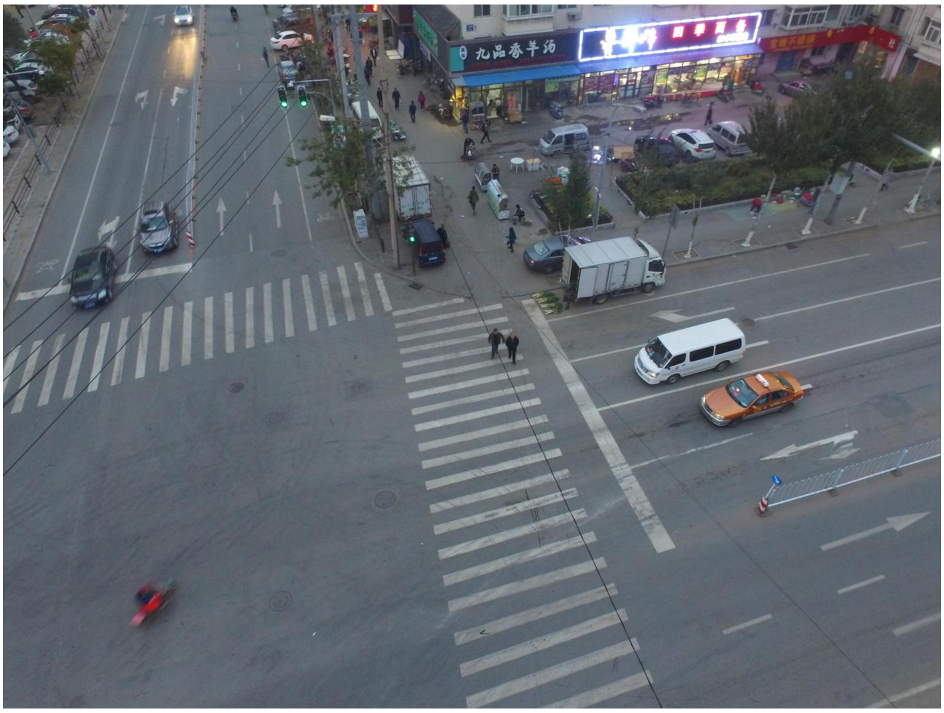


Figure A.8: Example image

Image shown in Figure A.8, outputs an answer like:

"The image captures a bustling street scene. Here’s a description of the objects and their arrangement:\n\n1. **Pedestrians**: On the sidewalk, several individuals are walking. They are scattered along the pavement, with some closer to the camera and others further away.\n\n2. **Vehicles**: The vehicles are scattered throughout the image, both moving and parked. On the left side of the image, there are multiple lanes dedicated to parked cars, with a mix of different brands and colors. In the middle of the image, there is a pedestrian crossing area marked by a white dashed line. On the right side, the lanes are clear with one moving white van being the most prominent.\n\n3. **Buildings**: There are buildings lining the street, with shops on the first floor and additional floors above them, possibly indicating residential housing. These buildings frame the scene in the background, providing context for the urban setting.\n\n4. **Cars in Motion**: Among the parked cars, there are a few cars that are in motion. One car is clearly visible in a position that indicates it's driving, moving toward the right side of the image.\n\n5. **Street Furniture**: There are traffic lights at various points along the street, regulating the flow of traffic.\n\n6. **Signage**: There is a neon sign with Chinese characters on it, possibly indicating a business or eatery, located on the right side of the image.\n\nThe objects in the image are arranged in a typical urban layout, with vehicles and pedestrians sharing the road, and buildings lining the sides. The scene is a common urban tableau, capturing the hustle and bustle of city life."

when prompted with Prompt No. 8:

`Describe the image with a focus on the arrangement of objects, such as their relative positions and orientations.`

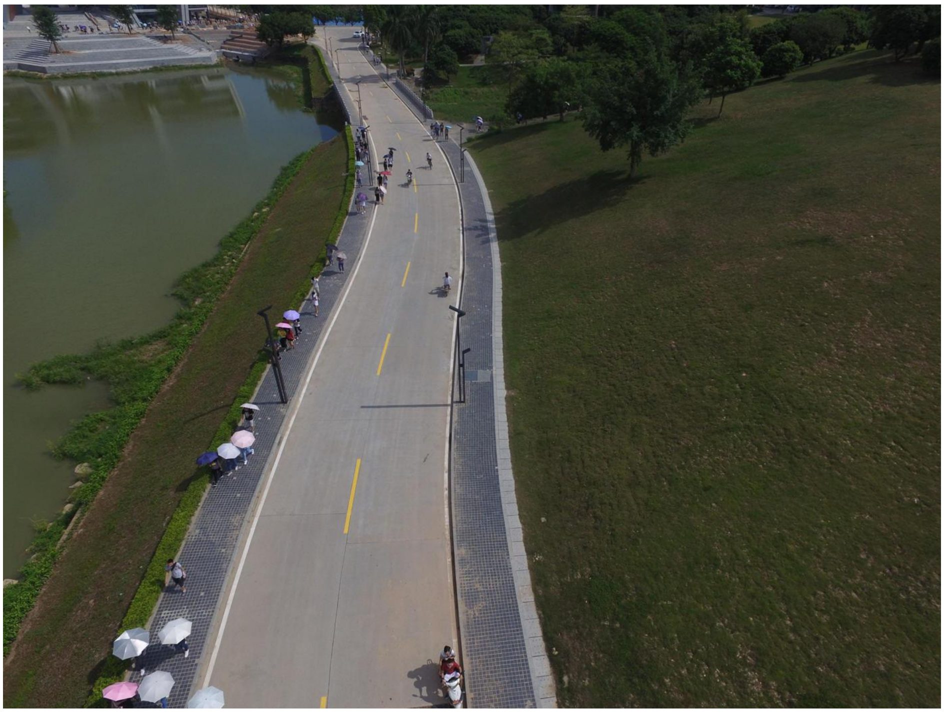


Figure A.9: Example image

Prompt No. 9 is:

`You are an expert in mapping and navigation with drones. Describe spatial relations in the given image taken by a drone.`

and an example result of it being prompted with image shown in Figure A.9 is:

```
"As an expert in mapping and navigation with drones, I can analyze the
spatial relationships between the various elements in the image:\n\n1.
Water Body: There is a body of water on the left side of the image.
The water appears to be stagnant, potentially a lake or a pond.\n\n2.
Divided Road: The image shows a long, paved road that is divided into two
segments, suggesting a separated lanes system or possibly a pedestrian
sidewalk adjacent to a separate lane for vehicles.\n\n3. People and
Bicycles: There are multiple individuals scattered along the pathway
near the water on the left. These people are likely pedestrians or
people riding bicycles, as suggested by the bicycles visible in the
image.\n\n4. Umbrellas and Seating: Closer to the waterfront, a series
of umbrellas and what appears to be seating or tables are set up,
suggesting an outdoor dining or relaxation area by the water.\n\n5.
Elevation Profile: As one looks down the road, the left side appears
to be higher in elevation than the right side, which is consistent
with the road crossing a raised natural barrier, such as an embankment
or a hill.\n\n6. Vegetation: There’s a green vegetation along the
divide between the road lanes, and a large expanse of green grass
on the right side of the image.\n\n7. Time of Day: The shadows of
the trees and the general lighting suggest that this image was taken
during the daytime, possibly midday given the soft shadows and lack
of harsh shadows.\n\n8. Pedestrian Path: The pathway adjacent to the
road has a distinct contrast in color compared to the road itself,
indicating that it is either paved with a different material or has
been treated differently to create a dedicated path for pedestrians or
cyclists.\n\nIn terms of navigation, the presence of the pathway, road
signs, and the general layout would allow a user to navigate this area
using maps from services like Google Maps, Apple Maps, or local mapping
applications. The umbrellas and seating suggest that this area could
be part of a commercial or recreational zone, so additional information
from businesses or local authorities could provide specific points of
interest and directions."
```