Spatial Information and Metadata Generation in Aerial Robotics with Vision-Language Models

Elena Wachtler*, Antonella Barišić Kulaš, Tamara Petrović, Stjepan Bogdan

***Abstract*— Unmanned aerial vehicles (UAVs) are used for a wide range of tasks, many of which are autonomous. Since UAVs operate in three-dimensional space, visual data presents a rich source of input for them. Meanwhile, the rapid development of artificial intelligence has given rise to vision-language models (VLMs). These models not only recognize objects in a scene, as traditional computer vision techniques do, but also demonstrate a strong understanding of scene context, enabled by large language models (LLMs) serving as one of their core components. However, applications as specialized as aerial robotics typically require custom-trained models, which demand significant computational resources, large datasets, and expensive hardware. Motivated by these limitations, we investigate whether a state-of-the-art, general-purpose VLM can be leveraged for UAV-relevant tasks through carefully designed prompts, without additional training. We evaluate the VLM's understanding of aerial scenes through the task of image captioning, using visual question answering that targets spatial information within the scene. Furthermore, we utilize VLMs to automatically generate metadata for aerial datasets, enabling more complex future tasks. Results show that VLMs can generate accurate metadata with high agreement with human-annotated data even without retraining. They also demonstrate that careful prompt engineering enhances the model's ability to reason about and attend to spatial information, highlighting its potential in aerial robotics applications.**

***Keywords*—aerial imagery, UAVs, spatial awareness, vision-language models**

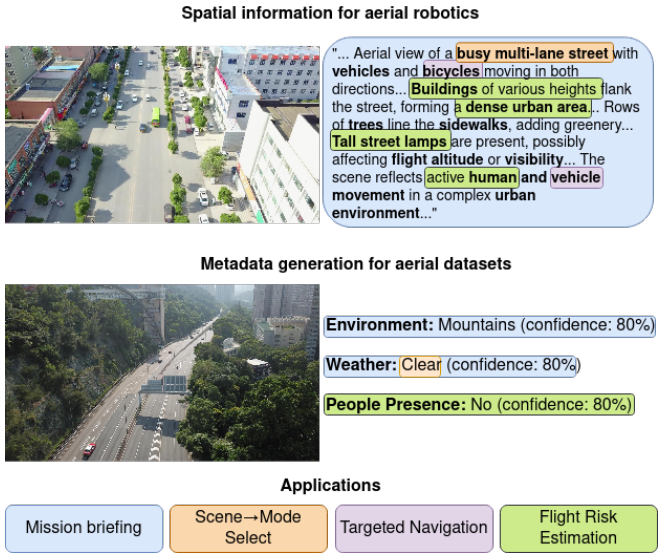

Fig. 1. Example outputs of VLMs applied to UAV imagery. The top image illustrates spatial scene understanding for aerial robotics, while the bottom shows automatic metadata generation for aerial datasets. These outputs enable downstream applications such as mission briefing, context-based module switching (e.g., selecting a different localization system in foggy conditions), airspace compliance, targeted navigation (e.g., follow green truck) and flight risk estimation. Best viewed in color.

I. Introduction

Vision-language models (VLMs) are a rapidly advancing field in artificial intelligence, combining visual and textual modalities to enable a deeper understanding of images [1]. While these models have shown promising results in tasks such as image captioning, object detection, and visual reasoning, their potential applications in robotics, particularly aerial robotics, remain largely unexplored. Aerial robotics presents unique challenges and opportunities due to complex spatial dynamics, diverse environmental conditions and the need for real-time decision making [2].

The richest source 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 general-purpose VLMs can be used in robotics, particularly given that training custom robotics VLMs is typically resource-intensive, requiring substantial computational resources, large specialized datasets and expensive hardware. Consequently, leveraging general-purpose VLMs through careful prompt engineering, without additional training, emerges as a compelling approach.

In this paper, we investigate the potential of state-of-the-art general-purpose VLMs for aerial robotics applications. Specifically, we examine whether these models can be leveraged through careful prompt engineering to interpret aerial imagery, extract spatial information and generate useful metadata as illustrated in Fig. 1. The figure demonstrates how VLMs can process complex aerial scenes to identify key environmental factors, weather conditions, human presence and spatial information. Such information is valuable for more advanced downstream tasks, including pilot briefing, targeted navigation, and flight risk assessment. Our findings provide insight into the design of prompts for extracting spatial information and demonstrate the potential of VLMs to enrich existing aerial data for more complex tasks. Our main contributions are:

1) An analysis of the impact of prompt engineering on the quality of robotics-related outputs, with special emphasis on extracting spatial information from aerial

*Corresponding author. E-mail: elena.wachtler@fer.unizg.hr.
All authors are with the Laboratory for Robotics and Intelligent Control Systems, University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia.

imagery.
2) A method for leveraging VLMs to automatically generate metadata for aerial robotics datasets, including weather conditions, environment types and the presence of people.
3) A comparative analysis evaluating the accuracy and reliability of VLM-generated metadata against human-labeled ground truth.

II. Related work

Vision-language models are a class 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]. These models go beyond traditional computer vision by not only recognizing objects in images, but also demonstrating an understanding of contextual relationships, and even the semantics of visual content, enabled in part by large language models being one of their main components. Recent advances include models such as LLaVA [4], Gemini [5], Flamingo [6] and Janus-Pro [7] among others. Despite rapid progress, VLMs continue to face challenges in computational efficiency, fine-grained understanding and spatial reasoning, along with concerns related to hallucination, safety and data scarcity.

VLMs in aerial robotics. Aerial imagery represents unique challenges such as object occlusion, low resolution, and the scarcity of images with the specific aerial angle in publicly available datasets [2]. However, the emergence of large pre-trained foundation models has introduced zero-shot and prompt-based capabilities [1] that offer broader generalization across tasks. CLIPSwarm [8] demonstrates this capability by leveraging the CLIP vision-language model to autonomously generate unmanned aerial vehicle (UAV) swarm formations from natural language prompts. Similarly, AirVista [9] fine-tunes LLaVA-1.5 on data synthesized from other VLM, depth estimator and segmentation model to enhance 3D spatial reasoning for UAVs. However, the efficacy of these methods can be constrained by factors such as the fidelity of synthetic data or the substantial computational overhead of fine-tuning. In broader robotics research, notable advances include the Gemini models [10]: Vision-Language-Action (VLA) model for direct robot control and Gemini Robotics-Embodied Reasoning (ER) focused on spatial and temporal understanding. In aerial robotics, VLMs are primarily applied to navigation and planning tasks, with their effectiveness dependent on their spatial reasoning skills. An important work AerialVLN [11] has recently published a large-scale benchmark for vision-and-language navigation tailored to UAVs operating in complex outdoor environments. Such specialized datasets and benchmarks are essential for addressing aerial robotics challenges like 3D spatial reasoning and language-guided flight control, as these capabilities must be tested on data that authentically represents UAV operational conditions. Another VLM-based approach for UAV navigation and object detection [12] utilizes nuisance-invariant feature extraction and transfer learning, but is evaluated on general-purpose datasets rather than UAV imagery. In contrast, our approach avoids the cost of model fine-tuning by emphasizing prompt engineering, metadata generation and analysis on real-world aerial data.

Spatial reasoning with VLMs. While VLMs generalize well across tasks, spatial reasoning remains a widely recognized limitation. Prior research shows that VLMs tend to focus on recognizing individual entities or general scene understanding rather than reasoning about spatial relationships between objects. For instance, [13] found spatial reasoning to be significantly more difficult for VLMs than identifying object properties, and [14] report performance worse than chance in spatial tasks. A growing consensus [15], [16], [17] positions spatial reasoning as a major bottleneck for current VLMs. Despite this, interest remains high across domains. Early efforts such as GQA [18] and synthetic datasets [19] offer structured spatial annotations, but typically in toy or indoor settings. More recent work uses scene graphs or annotated data for evaluation, either for general visual understanding [13], [20] or in robotics contexts like indoor navigation and manipulation [21], [22]. These efforts depend on datasets with dense spatial labels, which are lacking in real-world aerial scenarios. In aerial robotics, spatial tasks such as navigation, planning, and object localization have often been addressed using LLMs [23], [24], rather than VLMs. To the best of our knowledge, using VLMs to extract and evaluate spatial information from aerial imagery remains largely unaddressed. Furthermore, while datasets with known object positions in both synthetic and real space exist in other robotics scenarios, there is a notable lack of datasets that provide ground truth spatial information in realistic scenes of outdoor aerial scenery that would be relevant for aerial robotics. This gap motivates our investigation of whether spatial information can be extracted from real-world aerial imagery using a general-purpose VLM without fine-tuning, as detailed in the methodology and results that follow.

III. Methodology

A. Overview of the Approach

An overview of our approach is shown in Figure 2. We explore two tasks: (A) image captioning to extract spatial information and (B) metadata generation to enrich existing datasets. To this end, our pipeline begins by selecting the training set from the dataset and processing each image individually. For every image, the VLM is prompted with task-specific prompts. The model then generates textual responses, either as spatially descriptive captions (formulated as visual question answering) or structured metadata. Detailed descriptions of each task are provided in the following sections.

To enable real-time deployment on a UAV, we assume the use of 5G connectivity for offloading all VLM-related computation to a remote server, ensuring that no resource-intensive processing occurs on board the UAV.

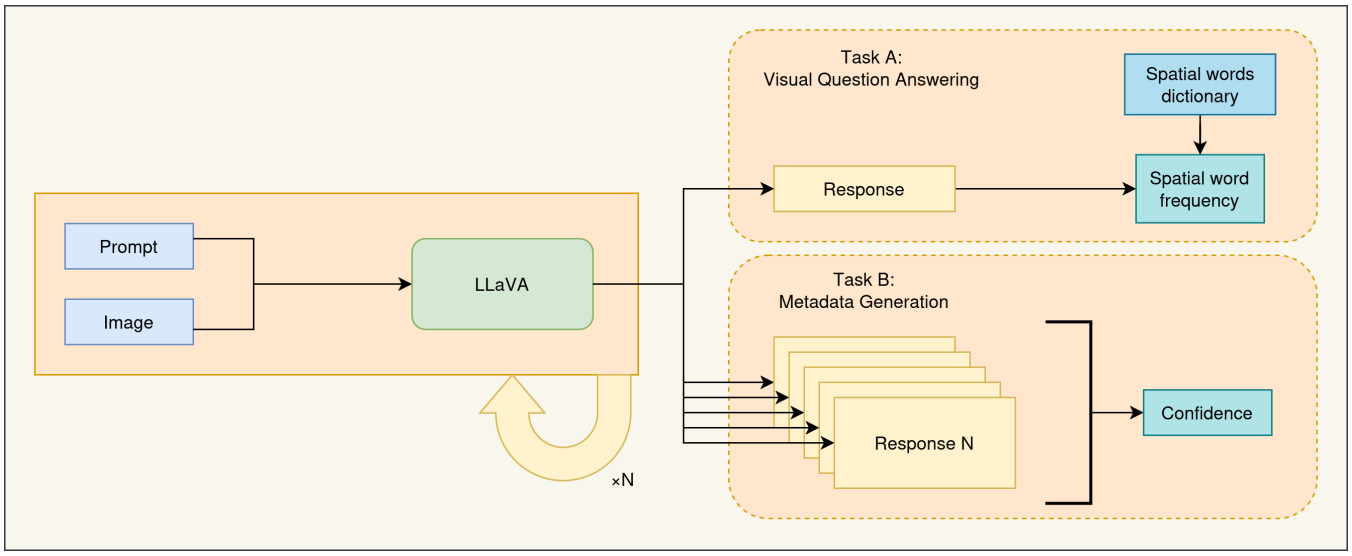

Fig. 2. Pipeline used for Visual Question Answering (Image Captioning) task: LLaVA is prompted for each (prompt, image) pair. The generated response is processed by counting spatial expressions from a predefined set (dictionary in the diagram) and calculating the spatial word frequency by dividing the count of found spatial expressions with the total number of words in the response. Pipeline for Metadata Generation task: LLaVA is prompted five times for each (prompt, image) pair. The five generated responses are then averaged, and the score is treated as confidence.

B. Image Captioning with VQA

Visual question answering (VQA) is a downstream task of correctly answering a question, given a visual input [25]. Unlike the usual classification task that would include choosing the most suitable answer from multiple options, in this case the model is asked for a free-form response, more similar to image captioning, but guided by a textual prompt. Since robots operate in three dimensional space, spatial information is of great importance to them. Hence, we decided to test the ability of vision-language models to spatially understand the scene. Three prompts request general information, testing the general focus of VLMs on spatial relations, while three remaining prompts targeted spatial information by giving direct cues to the VLM. Table I presents the prompts, which differ in focus, scope and length. Providing contextual information about the image content improves the performance of the model in understanding the scene [26]. To that end, prompts 1, 2, 3 and 6 provide contextual information emphasizing that the image was captured by an unmanned aerial vehicle, thereby encouraging the model to provide responses that are contextually relevant for UAV-based scene interpretation.

As mentioned earlier, for each of the images from the train set, a single response was generated. The response was then processed by counting the occurrences of spatial phrases from a predefined dictionary. The pipeline is outlined in Figure 2.

C. Metadata Generation

A single prompt was developed for the metadata generation task, aimed at extracting information relevant to unmanned aerial vehicle pilots. The prompt asked the vision-language model to label each image across three categories:

TABLE I

LIST OF PROMPTS USED FOR IMAGE CAPTIONING: GENERAL PROMPTS (1–3) AND SPATIAL PROMPTS (4–6)

Prompt #	Prompt Text
1	This image is an aerial photo taken by a UAV. Please provide detailed information about the image contents.
2	Describe the photo that shows an aerial view seen by a UAV, highlighting information that could be useful for the UAV.
3	Analyze the image from an aerial perspective captured by a UAV. Identify any areas of interest, including potential challenges for UAV navigation.
4	Describe the layout and spatial arrangement of the objects in the image.
5	Describe the image with a focus on the arrangement of objects, such as their relative positions and orientations.
6	You are an expert in mapping and navigation with UAVs. Describe spatial relations in the given image taken by a UAV.

- **environment**, to help evaluate terrain and potential hazards,
- **weather**, a critical factor for flight safety and mission planning,
- and **human presence**, important both for operational safety and privacy concerns.

Unlike the previous task of image captioning, the model was prompted with each image and the single prompt five times, resulting in five generated responses, as shown in Figure 2. Consistent outputs across multiple prompt runs, even with sampling enabled, emerged as a notable strength of the model. Rather than variability, the model often converged on the

same response, signaling a high degree of certainty and stability in its predictions. This is the prompt presented to the model:

Prompt

You are a UAV 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 / 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 prompt is deliberately precisely structured to specify both the expected format and the predefined labels for each category. It underscores the need to assign a label to every category, thereby reducing the likelihood of post-processing issues related to formatting errors or missing data.

D. Ground Truth Data Collection

To evaluate metadata generated by the VLM, three individuals from diverse professional backgrounds annotated a randomly selected sample of 200 images from the VisDrone dataset [27], which served as ground truth. Specifically, for each of the images from the sample, they were given three multiple choice questions, in which they were asked 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.

IV. Results

A. Implementation details

Due to the limited accessibility of many commercial vision-language models, we opted for an open-source state-of-the-art model to ensure reproducibility. LLaVA (**L**arge **L**anguage-**a**nd-**V**ision **A**ssistant), a multimodal model that integrates a vision encoder with a large language model, was selected for its strong performance and open availability [28]. We used LLaVA-NeXT (LLaVA-1.6), specifically its `llava-v1.6-mistral-7b-hf` checkpoint, available on HuggingFace [4]. As the name suggests, the checkpoint uses Mistral-7B [29] as its language model. While we acknowledge the value of comparing across multiple VLMs, we argue that LLaVA-NeXT is representative of modern VLM capabilities and sufficiently expressive for the studied tasks. In our setting, the goal is not to benchmark VLMs per se, but to probe the feasibility of adapting a general-purpose VLM for specialized aerial tasks through prompt engineering. We therefore prioritize depth of analysis using a single, well-supported model over breadth.

Since every checkpoint is trained with a specific prompt format, depending on the underlying large language model backbone [30], the corresponding prompt structure was maintained during prompting: `"[INST] <image>\nWhat is shown in this image? [/INST]"`, where the prompt text is substituted appropriately. We set the maximum number of output tokens denoted by t_{max} to 500 and enabled sampling to ensure response diversity across runs.

B. Dataset

Given the challenges described in Section I, we decided to test the generalizability of a general-purpose model by measuring its effectiveness on aerial imagery. For this work, a publicly available dataset was employed, containing diverse visual scenarios captured by UAVs, making it suitable for tasks requiring an understanding of spatial relationships and scene interpretation, such as UAV navigation, surveillance, disaster response, and similar.

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]. VisDrone benchmark dataset consists of over 280 video clips formed by over 260,000 frames and over 10,000 static high-resolution images, captured by various cameras mounted to UAVs that varied in their model. The images were taken in fourteen Chinese cities, capturing different objects, scenes with density ranging from sparse to crowded, where the weather and lighting conditions also varied. It is divided into training (6,471 image), validation (548 images) and testing (1,580 images) datasets. In this work, only the training dataset images were used, while the accompanying annotations were discarded.

C. Evaluation Methods

Two approaches were used to evaluate image captioning: a qualitative analysis and the quantitative measure of spatial phrase frequency. To conduct the qualitative analysis, a subset of VLM responses was manually reviewed to assess the relevance, spatial awareness and contextual grounding. For quantitative evaluation, we define spatial phrase frequency (SF) as the proportion of phrases that describe spatial relationships (e. g., "next to," "above," "in front of") relative to the total word count:

$$SF = \frac{N_s}{N_t}, \tag{1}$$

where N_s is the number of spatial phrases and N_t is the total number of words in the response.

To evaluate metadata generation, we define accuracy (A) as a ratio of correctly predicted category labels to the total number of samples:

$$A = \frac{N_c}{N}, \tag{2}$$

where N_c is number of correct labels in a certain category and N is the total number of images. The generated responses were filtered so that only those responses which provide a valid label from the predefined set of category labels were considered. If provided, additional descriptions and reasoning of label choice were ignored, and only the category labels were extracted for further quantitative evaluation.

D. Image Captioning Results

Fig. 3. Part of VLM's response to Prompt 5 paired with image shown above, which demonstrates the model's ability to understand the semantics of image content (underlined):
"(...) On the far right, there's a green umbrella, possibly belonging to a street vendor or an urban decoration. (...)"

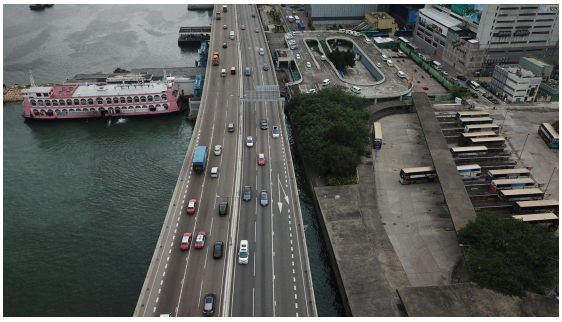

Fig. 4. An example response to Prompt 3 paired with image above:
"From an aerial perspective, the image shows a busy urban area with a focus on a major *bridge crossing over a body of water*. The bridge is flanked by *streets on either side*, and there are various modes of transportation visible, including cars, buses, and trucks. A *ferry is docked alongside the bridge on the left side*, suggesting this might be a transit hub.
In terms of potential challenges for UAV navigation:
1. **Airspace Restrictions**: The proximity of the bridge and the urban environment might make flying a UAV in this area subject to airspace restrictions or regulations, which could limit one's ability to navigate freely.
2. **Dense Traffic**: The extensive use of ground-level vehicles could lead to increased risks related to UAV navigation. Dense vehicle traffic might also obstruct the line of sight and require careful planning to avoid collisions.
3. **Nearby Waterways**: Operating a UAV *near water* can be challenging due to the possibility of wind gusts off the water surface, which could disrupt its flight path, and the fact that water is a significant obstacle in case of any equipment malfunction or loss of control. (...)"

The qualitative results are illustrated across two figures. Figure 3 shows an image and a segment of the caption given as response to prompt 5. Aside from accurately identifying objects (*green umbrella*), the model displays a deeper understanding of the content semantics (underlined), proposing the object's likely function.

Another example of a response can be seen in Figure 4, with underlined segments again indicating moments of contextual and content understanding. Moreover, since this response was obtained by a prompt directed to spatial information, the acquired spatial information are marked in italics. These examples illustrate the model's capability to provide valuable answers in accordance with the prompt, along with its competence in grasping the image content, as well as content meaning.

To measure the spatial information contained in the 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.

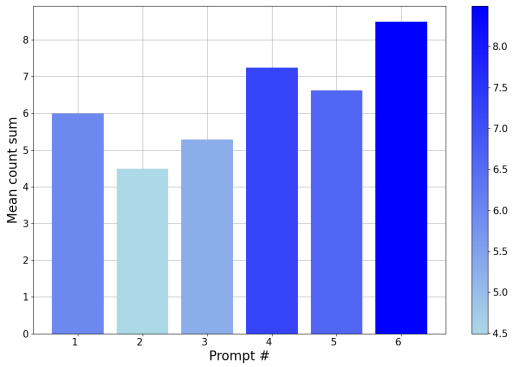

Fig. 5. Count of spatial phrases in responses for each prompt for VisDrone dataset

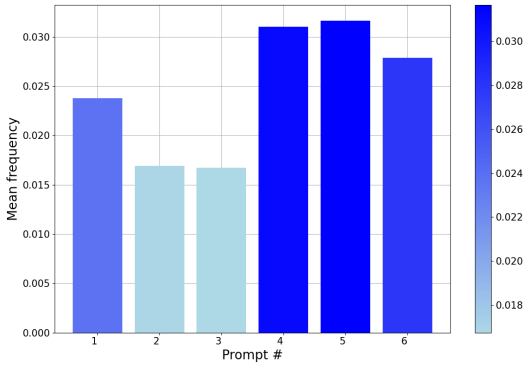

Fig. 6. Frequency of spatial phrases in responses for each prompt for VisDrone dataset

The number of spatial phrases in the VLM's responses to each of the six prompts is shown in Figure 5. However, since some responses are longer than others, raw count may not accurately reflect the proportion of spatial information. To address this, Figure 6 presents the frequency of spatial phrases in the response, taking the length of the response into account. We observe that prompts specifically designed to elicit spatial reasoning (prompts 4–6) induced responses with a substantially higher density of spatial content.

Furthermore, despite not explicitly requesting spatial information, prompt 1 still yielded a high degree of content describing spatial layout. This is likely caused by the focus of the prompt on image content — due to the aerial perspective, such descriptions naturally involve spatial relationships among the objects depicted.

Although there are slight differences in spatial phrase frequency between the spatial prompts, these differences are not deemed substantial enough to be considered significant.

The top word cloud in Figure 7 illustrates the most frequent terms found across all collected responses. Knowing that the dataset images were largely captured in urban environments, it is not surprising that terms such as *building*, *road*, *street*, or *vehicle* appear frequently. In parallel, the bottom word cloud in Figure 7 shows the occurrence distribution of predefined spatial phrases in the responses. For the sake of legibility, words *by* and *along*, despite being part of the predefined spatial phrases dictionary, were excluded from the word cloud, as they can appear in grammatical constructions that are not related to spatial meaning as well.

Fig. 7. Word clouds in the image captioning task: top shows the frequency of all words; bottom shows the frequency of only predefined spatial words.

E. Generated Metadata Results

The distribution of labels that were generated by the model is presented in Table II. As expected for the VisDrone dataset, in the *environment* category, *city* is a highly dominant label, with *mountains* and *sea* only just appearing. Furthermore, most of the images are labeled as *clear* weather, then *nighttime* and *overcast*, with only a very small part of the images marked as foggy or rainy weather. Given that the images from the dataset were taken by UAVs that have difficulty flying in fog or rain, this also aligns with expectations.

TABLE II
DISTRIBUTION OF LABEL PERCENTAGES ACROSS CATEGORIES

Environment			
City	95.54%	Rural	2.67%
Forest	0.91%	Desert	0.75%
Sea	0.09%	Mountains	0.04%
Weather			
Clear	76.66%	Overcast	5.94%
Nighttime	17.13%	Foggy	0.18%
Rain	0.09%		
Humans			
Yes	77.28%	No	22.72%

Consistent with the methodology employed in processing human annotated metadata, from the five responses generated by LLaVA, the label that was selected in most responses was taken as the final answer. Ambiguous answers that contained multiple or invalid labels for a category were ignored. Draws were handled by selecting the first provided answer as final. The category *environment* has proven most successful, with an accuracy of 93%. The category *weather* follows second at 87.5%, while the category *humans* was lowest with 69.5%, as shown in Table III. To conclude, the categories *environment* and *weather* both showed high accuracy, while we believe that ambiguities in determining human presence in the scene caused the lower accuracy in determining human presence.

TABLE III
ACCURACY OF DIFFERENT CATEGORIES COMPARED TO HUMAN ANNOTATED IMAGES

Category	**Accuracy (%)**
Environment	93
Weather	87.5
Humans	69.5

For instance, the leftmost image in Figure 8 is an example when annotators and the model disagreed on the label *humans*. In four out of five responses, the model said there are humans present. Furthermore, in two responses, it offered an explanation for this choice — there have to be humans driving the moving cars: `(...) humans: yes (vehicles driving on the freeway)` and `(...) humans: yes (there are vehicles visible)`. All human annotators labeled this image as no human presence. This demonstrates that the VLM understands that even though people are not visible, they are in the cars in motion. We interpret these disagreements as application-dependent semantic differences: while in applications such as traffic monitoring or emergency response, moving vehicles may

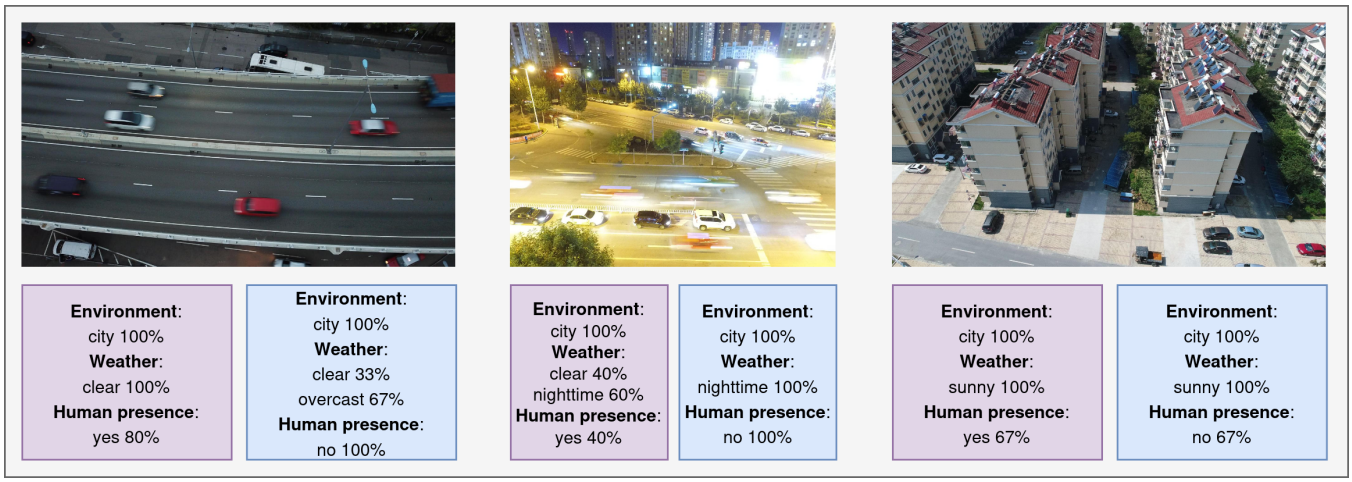

Fig. 8. Examples of metadata disagreement between the model and human annotators. Purple boxes (left) represent generated metadata categories that the model provided; blue boxes (right) show human annotators' choice.

reasonably imply human presence, tasks such as pedestrian detection or privacy-sensitive analysis may require only directly visible humans to be considered. Importantly, the behavior of VLMs in such ambiguous scenarios can be guided through prompt design or explicit label definitions, enabling adaptation to different operational requirements. The given example therefore helps clarify what could be the cause of lower accuracy observed in determining human presence compared to the other two categories.

In the initial metadata generation configuration, the weather category had *sunny* as a label option as well. However, after the labeling process, some annotators reported difficulty distinguishing between labels *clear* and *sunny*, so these two labels were merged to *clear*, and the model prompt that was used for the whole VisDrone dataset only offered this option.

Examples of images where the labels provided by the model and the annotators differed are shown in Figure 8. These examples illustrate that disagreements often come from ambiguity in the image content, where more than one label could be considered valid.

F. Discussion and Interpretation of Findings

Results have proven the capability of the model to understand the scene context, elevating pure object detection to reasoning about the function of the object in a scene. The model generated responses that were relevant in the vast majority of cases, but irrelevant information does appear. Calculating spatial phrase frequency proved the positive influence prompt engineering can have on spatial content in responses. Even though label ambiguity affected how certain categories were assigned, the model still achieved high accuracy in generating metadata compared to human annotations. On top of that, the model exhibited human-analogous performance, being conflicted between labels similar to human annotators.

V. Conclusion and Future Work

This work has demonstrated that, even without domain-specific adaptation, the chosen general-purpose VLM, LLaVA-1.6, can produce valuable textual descriptions of aerial imagery. The responses illustrated the reasoning capabilities of the VLM, showing understanding of meaning and purpose of recognized objects. Furthermore, the quantitative analysis of image captioning conducted through a visual question answering task has demonstrated the favor of the model to increase the quantity of the spatial information in a response, when prompted accordingly. However, due to the lack of ground truth information and the complexity of evaluating the accuracy of such content, the measure of spatial phrase frequency that was used did not account for the accuracy of the provided spatial information. To address this, future work could focus on constructing datasets specific to aerial robotics with annotated ground truth for spatial relationships, enabling a precise evaluation of spatial accuracy. In addition, future efforts could explore automated prompt generation to replace hand-crafted prompts, enabling seamless integration into autonomous robotics systems.

Generating metadata from aerial images was also proven successful, with the results provided by the model closely aligning with human annotations. The VisDrone dataset, with its real-world, outdoor imagery, was particularly useful in this regard, as it provided valuable data relevant for aerial robotics applications. Unlike controlled lab settings, VisDrone reflects the complexities of urban environments, offering a more realistic representation of what VLMs may encounter in operational scenarios. That being said, the dataset is primarily focused on urban city environments, and as such, certain category labels are underrepresented. Future work would benefit from expanding the diversity of the dataset, incorporating a wider range of environments to better generalize the model's capabilities. While we recognize that having more diverse environments would provide a more comprehensive understanding, to our knowledge, no large-

scale aerial imagery datasets currently include such diversity across various environments, which presents an opportunity for future data collection.

In conclusion, this study provides valuable insights into the performance of a vision-language model on aerial imagery. The model demonstrated promise in recognizing objects and reasoning about the environment. Future work should include prompts that test for a wider range of applications, explore more diverse datasets, whether synthetic or real, and enhance evaluation metrics.

ACKNOWLEDGMENT

This work was in part supported by the ASTRA Project (An advanced system for the automatic generation of synthetic thermal and color images with highly precise annotations) funded by the European Union - NextGeneration EU under grant NPOO.C3.2.R3-I1.05.0226. Additional support was provided through the project "POmorski Siguran autonomni sustav za Ekologiju, IDentifikaciju, Očuvanje i Nadzor zaštićenih područja (POSEIDON)" at the University of Zagreb Faculty of Electrical Engineering and Computing, funded by the EU - NextGeneration.

REFERENCES

[1] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, p. 5625–5644, Aug. 2024.

[2] D. Cazzato, C. Cimarelli, J. L. Sanchez-Lopez, H. Voos, and M. Leo, "A survey of computer vision methods for 2D object detection from unmanned aerial vehicles," *Journal of Imaging*, vol. 6, no. 8, p. 78, Aug. 2020.

[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] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "LLaVA v1.6 Mistral-7B," 2023. [Online]. Available: https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf

[5] Gemini Team et al., "Gemini: A family of highly capable multimodal models," 2023. [Online]. Available: https://arxiv.org/abs/2312.11805

[6] Alayrac et al., "Flamingo: a visual language model for few-shot learning," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 23 716–23 736.

[7] X. Chen, Z. Wu, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, and C. Ruan, "Janus-Pro: Unified multimodal understanding and generation with data and model scaling," 2025. [Online]. Available: https://arxiv.org/abs/2501.17811

[8] P. Pueyo, E. Montijano, A. C. Murillo, and M. Schwager, "CLIP-Swarm: Generating drone shows from text prompts with vision-language models," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Oct. 2024, p. 11917–11923.

[9] F. Lin, Y. Tian, Y. Wang, T. Zhang, X. Zhang, and F.-Y. Wang, "AirVista: Empowering UAVs with 3D spatial reasoning abilities through a multimodal large language model agent," in *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, Sep. 2024, p. 476–481.

[10] Gemini Robotics Team et al., "Gemini robotics: Bringing AI into the physical world," 2025. [Online]. Available: https://arxiv.org/abs/2503.20020

[11] S. Liu, H. Zhang, Y. Qi, P. Wang, Y. Zhang, and Q. Wu, "Aerialvln: Vision-and-language navigation for UAVs," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 15 384–15 394.

[12] Y. Li, L. Yang, M. Yang, F. Yan, T. Liu, C. Guo, and R. Chen, "NavBLIP: a visual-language model for enhancing unmanned aerial vehicles navigation and object detection," *Frontiers in Neurorobotics*, vol. 18, Jan. 2025.

[13] F. Liu, G. Emerson, and N. Collier, "Visual spatial reasoning," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 635–651, 06 2023.

[14] J. Wang, Y. Ming, Z. Shi, V. Vineet, X. Wang, Y. Li, and N. Joshi, "Is a picture worth a thousand words? delving into spatial reasoning for vision language models," 2024. [Online]. Available: https://arxiv.org/abs/2406.14852

[15] I. Stogiannidis, S. McDonagh, and S. A. Tsaftaris, "Mind the gap: Benchmarking spatial reasoning in vision-language models," 2025. [Online]. Available: https://arxiv.org/abs/2503.19707

[16] A. Kamath, J. Hessel, and K.-W. Chang, "What's "up" with vision-language models? investigating their struggle with spatial reasoning," 2023. [Online]. Available: https://arxiv.org/abs/2310.19785

[17] S. Chen, T. Zhu, R. Zhou, J. Zhang, S. Gao, J. C. Niebles, M. Geva, J. He, J. Wu, and M. Li, "Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas," 2025. [Online]. Available: https://arxiv.org/abs/2503.01773

[18] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," 2019. [Online]. Available: https://arxiv.org/abs/1902.09506

[19] R. Liu, C. Liu, Y. Bai, and A. L. Yuille, "Clevr-ref+: Diagnosing visual reasoning with referring expressions," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4180–4189.

[20] F. Shiri, X.-Y. Guo, M. G. Far, X. Yu, G. Haffari, and Y.-F. Li, "An empirical analysis on spatial reasoning capabilities of large multimodal models," 2024. [Online]. Available: https://arxiv.org/abs/2411.06048

[21] W. Cai, I. Ponomarenko, J. Yuan, X. Li, W. Yang, H. Dong, and B. Zhao, "Spatialbot: Precise spatial understanding with vision language models," 2025. [Online]. Available: https://arxiv.org/abs/2406.13642

[22] C. H. Song, V. Blukis, J. Tremblay, S. Tyree, Y. Su, and S. Birchfield, "Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 15 768–15 780.

[23] W. Wang, Y. Li, L. Jiao, and J. Yuan, "Gsce: A prompt framework with enhanced reasoning for reliable llm-driven drone control," 2025. [Online]. Available: https://arxiv.org/abs/2502.12531

[24] A. Tagliabue, K. Kondo, T. Zhao, M. Peterson, C. T. Tewari, and J. P. How, "Real: Resilience and adaptation using large language models on autonomous aerial robots," 2023. [Online]. Available: https://arxiv.org/abs/2311.01403

[25] 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.

[26] X. Wang and Z. Zhu, "Context understanding in computer vision: A survey," *Computer Vision and Image Understanding*, vol. 229, p. 103646, Mar. 2023.

[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] 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/

[29] 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

[30] 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