Large Models in Dialogue for Active Perception and Anomaly Detection

Tzoulio Chamiti[1], Nikolaos Passalis[2], and Anastasios Tefas[1]

[1] Computational Intelligence and Deep Learning Group, AIIA Lab., Dept. of Informatics
[2] Dept. of Chemical Engineering
Aristotle University of Thessaloniki, Thessaloniki 541 24, Greece
t.chamiti@csd.auth.gr, passalis@auth.gr, tefas@csd.auth.gr

Abstract. Autonomous aerial monitoring is an important task aimed at gathering information from areas that may not be easily accessible by humans. At the same time, this task often requires recognizing anomalies from a significant distance and/or not previously encountered in the past. In this paper, we propose a novel framework that leverages the advanced capabilities provided by Large Language Models (LLMs) to actively collect information and perform anomaly detection in novel scenes. To this end, we propose an LLM-based model dialogue approach, in which two deep learning models engage in a dialogue to actively control a drone to increase perception and anomaly detection accuracy. We conduct our experiments in a high fidelity simulation environment where an LLM is provided with a predetermined set of natural language movement commands mapped into executable code functions. Additionally, we deploy a multimodal Visual Question Answering (VQA) model charged with the task of visual question answering and captioning. By engaging the two models in conversation, the LLM asks exploratory questions while simultaneously flying a drone into different parts of the scene, providing a novel way to implement active perception. By leveraging LLM's reasoning ability, we output an improved detailed description of the scene going beyond existing static perception approaches. In addition to information gathering, our approach is utilized for anomaly detection and our results demonstrate the proposed method's effectiveness in informing and alerting about potential hazards.

Keywords: Active Anomaly Detection · LLM · VQA · Aerial Monitoring

1 Introduction

In the last few years, drones have witnessed numerous technological advancements, as well as great commercial exposure for their ability to perform difficult tasks, such as surveillance, anomaly detection, and aerial monitoring in challenging environments. To effectively support these tasks and ensure the efficient and autonomous operation of robots, large informative datasets, e.g., containing drone images, action states, and/or anomalies, were necessary in order to

cover every possible scenario that could occur [1–3]. These approaches primarily focused on collecting a large quantity of data and employing different learning techniques to detect possible anomalies in autonomous drone flying scenarios.

With the major advancements in deep learning across numerous domains, there have been multiple attempts to incorporate these modern, more effective technologies for the sake of enhancing autonomous systems' efficiency and capability. By deploying larger, more advanced deep learning models a substantial improvement in performance was witnessed [4, 5]. Nevertheless, these methods lack the ability to *actively perceive* the scene in order to issue the appropriate control commands and further improve the perception accuracy based on the current conditions. Such active perception approaches have shown promising results in other relevant domains in recent years [6–8]. However, it is not trivial to implement such methods in open-world setups.

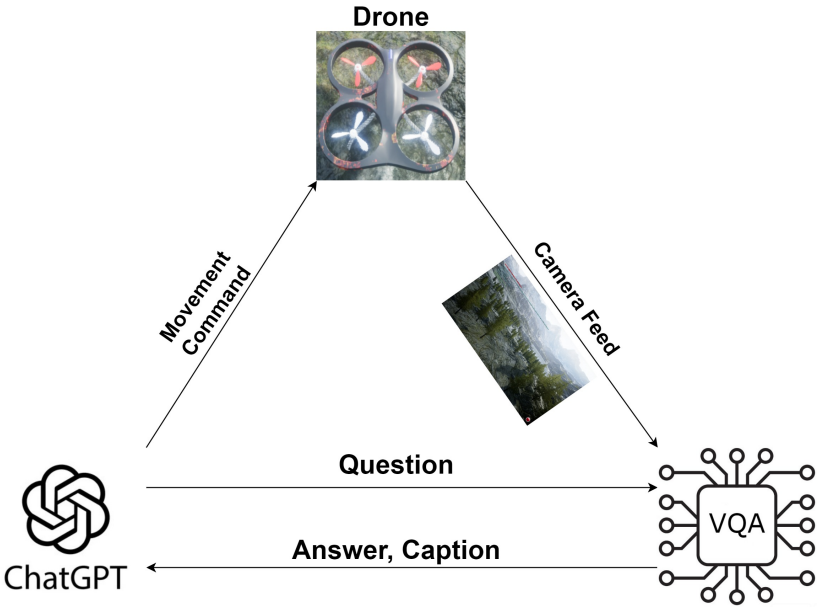

Fig. 1: Overview of the proposed model dialogue approach. First a drone captures an image. This image, along with an appropriate question, is fed to the employed VQA model. Then, the VQA model provides a response that is fed to the LLM model which in turn issues a movement command and a new exploratory question.

The main contribution of this paper is a novel approach for active perception and anomaly detection that leverages the capabilities of recent Large Language Models (LLMs) by developing a *model dialogue* approach in which two deep learning models interact in order to continuously improve the final pre-

diction. To this end, we equip the employed LLM with complete navigational control through a set of specific textual commands that ultimately navigate a drone in real time, implementing an active perception scheme in which the drone explores the scene and exploits potential hazardous scenarios and anomalies. Furthermore, we incorporate a Visual Question Answering (VQA) model in order to engage the two models in interactive conversation from which the LLM acts as a controller that can extract meaningful textual information about the unknown scene in which the drone operates. Our goal is to provide a detailed description of the scene gathered throughout the conversation along with explanations that led to these decisions. This dialogue process leads to an active perception pipeline in which we can gather additional information about the scene, as well as validate the scene details. The conducted experimental evaluation shows that the proposed approach can indeed enable a drone to successfully navigate an unknown open environment and provide an explainable and detailed description of the scene in a zero-shot fashion, as well as detect anomalies and output potential safety measures in response to potentially hazardous observations. The code used for the conducted experiments, including detailed prompts and experimental results are provided at `https://github.com/Tzoulio/Large_Models_Dialogue_for_Active_Perception`.

The rest of the paper is structured as follows. Section 2 introduces the related work, while the proposed method is presented in Section 3. The experimental evaluation is provided in Section 4, while Section 5 concludes the paper.

2 Related Work

The task of Visual Question Answering [9] has increased in popularity in recent years, with the ability to combine computer vision with Natural Language Processing (NLP) resulting in a system that can process two types of different modalities at the same time. Such an ability is crucial in robotics applications considering they are often applied to scenarios and environments that require handling such multimodal data. By giving a robot the ability to process multiple data together at once, they increase the quality and quantity of information they acquire, which in turn expands their overall knowledge of the world. As a result, there have been multiple attempts at applying VQA in robotics. Some works focus on having the robot interact with the environment and come up with an answer to a specific question, mimicking the VQA task. Deng et al. [10] uses VQA in a robotic manipulation scenario. They train a Deep Q Network (DQN) and through reinforcement learning teach the robot to continuously manipulate objects until they come up with the right answer. In [11] a Hierarchical Interactive Memory Network (HIMN) was deployed as a controller that allows the system to store and retrieve information hierarchically in the form of memory and enables the robot to provide an answer by interacting with its environment in real-time. EmbodiedQA [12] is another approach that deploys a robot in an unknown environment in which the robot learns to navigate through using imitation learning and ultimately gathers the appropriate information to answer the

question. Our work leverages the recent advances in VQA as a fundamental part of the proposed pipeline by employing a VQA model which acts as the *sensing* model, which processes the data acquired from the world and answers questions regarding these.

After the breakthrough that LLMs made in the field of AI, researchers have been constantly finding ways to utilize them in robotic applications. A lot of works leverage the LLMs' reasoning capabilities and language understanding ability to act as a communicator between the human operator who issues a command in natural language and the robot who executes the command in the form of code [13–16]. These approaches either directly map specific commands to code snippets that are applied on the robot directly or provide enough resources to the LLM to construct code and make specific API calls that will produce the correct result on the robot, as specified in the natural language prompt. Generally, a lot of research is focused on advancing the LLM capabilities further, by implementing different modules together with the LLM in an attempt to give it multi-modal capabilities [17–19]. This resulted in a lot of works which combined multi-modal variations of LLMs into robot task planning [20–22]. These works utilize imitation learning to teach a control agent how to perform the natural language tasks which are learned from a dataset consisting of sets of demonstrations during different timestamps. In other works, such as [13], users are able to control an aerial drone through natural language and prompt engineering. The proposed method goes beyond these approaches by employing a dialogue-based approach, in which only one model has full access to the visual modality and the other model can interact with this model through textual prompts.

The proposed method is more closely related to recent attempts to combine LLMs with VQA models. Some works [23–26] focus on initiating a conversation between the two models to enhance the VQAs ability in the captioning task. They start with a general caption of a query image and through ChatGPT's ability of understanding and generalising textual information an active dialogue between the LLM and the VQA module is initiated. During the dialogue, ChatGPT makes inquiries about possible information that the image might contain. Afterwards, the VQA model answers by confirming or denying and providing additional information for the scene. The process continues until ChatGPT outputs a detailed description containing all the knowledge it gathered through the conversation. Other methods follow a similar approach [27–29] by providing complementary knowledge to the LLM in the form of captions. This enhances the quality and flow of information, resulting in better answers and captions for the query images. Our method builds on this idea, going beyond these approaches by implementing *active perception* through the drone's navigation scheme. We collect a different image of the scene each time the drone reaches a new position. At the same time, the employed LLM asks an exploratory question with each movement command and the VQA model provides an answer and a caption. This way, we are able to gather more information (extracted by the different captions we get in every position) as well as explore parts of the initial image

that the camera could not see either by them being obscured or simply by being too far away.

3 Proposed Method

In this work, we aim to equip a drone with active perception and anomaly detection capabilities in order to provide a robust scene description, as depicted in Fig. 1. First, the drone leverages a VQA model which provides descriptions of the environment through captions. In this way, the VQA model provides a way for the LLM to "sense" the environment through text. Additionally, the VQA model outputs an image-caption matching score in order to help the LLM distinguish between good and bad captions. Then, the LLM validates the gathered textual information through the VQAs question-answering module combined with active perception and ultimately provides a generalized scene description together with explainable attention maps. The outline of the proposed approach is shown through an example in Fig. 2. This example should be used as a reference point through the description provided in this Section, since it further clarifies how the proposed method works.

For the VQA model, we incorporate the Plug-and-Play VQA (PnP-VQA) [30] framework, as shown in Fig. 3. To perform the task of image captioning, image-question pairs are processed by a pre-trained vision-language model called BLIP [31] which is also able to output a similarity score between the image and the question. The image is split into K patches and through GradCAM [32], a feature-attribution interpretability technique, they are able to provide the most relevant image patches. Finally, the image captioning module of BLIP is combined with top-k sampling to generate captions only for the relevant patches. Subsequently, the produced caption and question are fed into the question answering module to produce the answer. For the LLM, we employed the GPT3.5 as our model [33].

Let the LLM model denoted by $f(\mathbf{A}, \mathbf{C})$, which takes two distinct text sequences as input $\mathbf{A} = [A_1, A_2, \ldots, A_n]$, $\mathbf{C} = [C_1, C_2, \ldots, C_m]$ and outputs a response sequence $\mathbf{Q} = [Q_1, Q_2, \ldots, Q_k]$, in the form of a question i.e., $\mathbf{Q} = f(\mathbf{A}, \mathbf{C})$, where $\mathbf{A}$ denotes the answer to a previous question by the VQA model (if exists) and $\mathbf{C}$ denotes a textual description (caption) of the current scene. In this work, we employed the GPT3.5 model to implement $f(\cdot)$, while we feed the concatenated $\mathbf{A}$ and $\mathbf{C}$ to the model. We assume A_i, C_i and Q_i denote the indices of words, while n, m and k denote the corresponding sequence lengths. Similarly, the VQA network $g(\mathbf{Q}, \mathbf{I})$ takes as input the output sequence of the LLM $\mathbf{Q}$, as well as an image $\mathbf{I}$, producing two different textual sequences $\mathbf{A}, \mathbf{C} = g(\mathbf{Q}, \mathbf{I})$, where $\mathbf{A}$ is the answer to the question and $\mathbf{C}$ denotes the caption for the image. Then, these outputs are fed to the LLM and this process repeats in an iterative fashion.

To grant the LLM control of the drone we first define a set of diverse functions, each one in charge of a specific navigational output. Afterwards, we provide the drone with a detailed prompt consisting of a set of commands mapped to

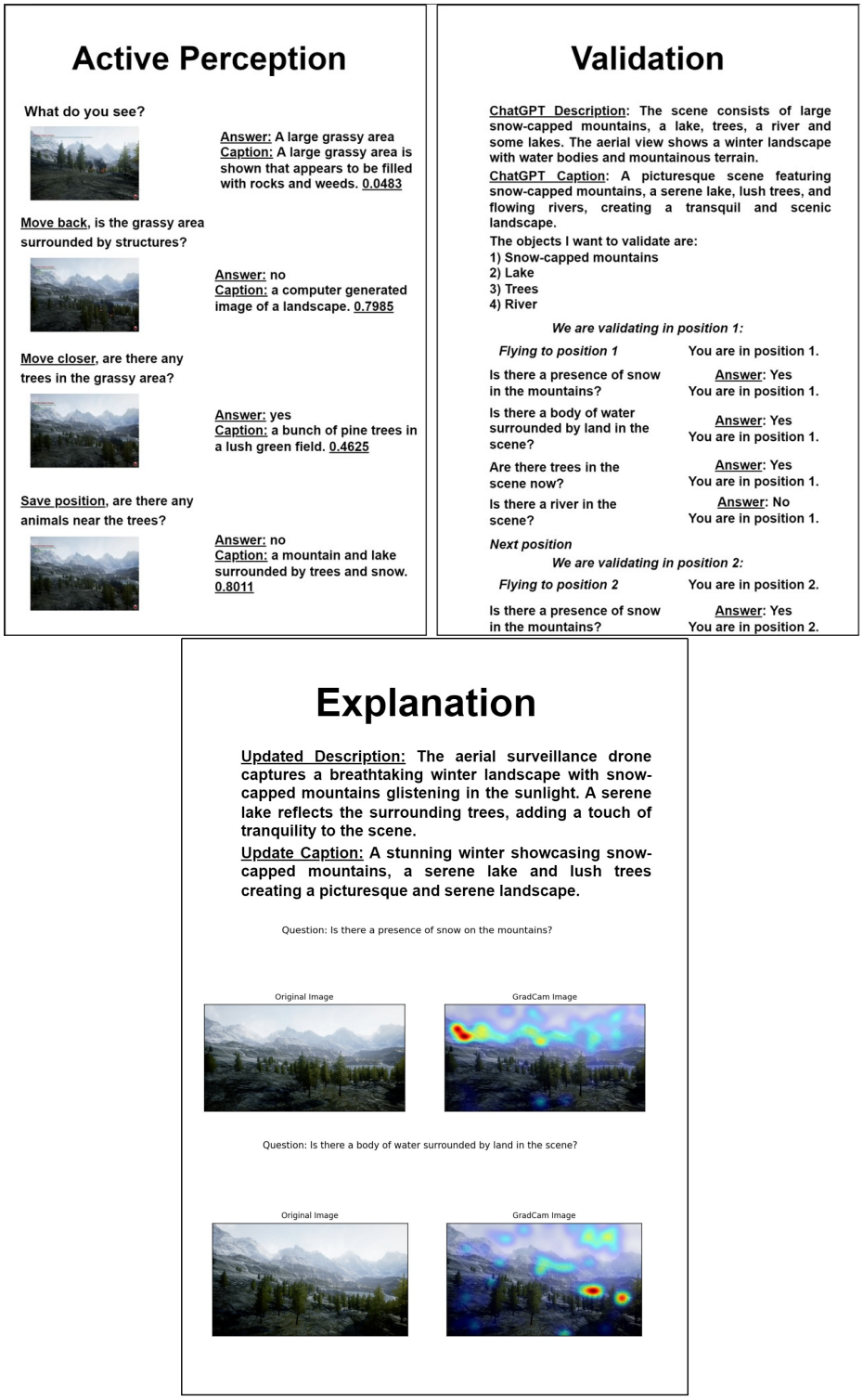

Fig. 2: A typical example of the operation of the proposed method. During active perception, the two models engage in a conversation and exchange information. In validation, a premature description and caption are chosen together and information is validated by revisiting the saved positions. Then, in the explanation mode, the final description and caption are provided together with attention maps.

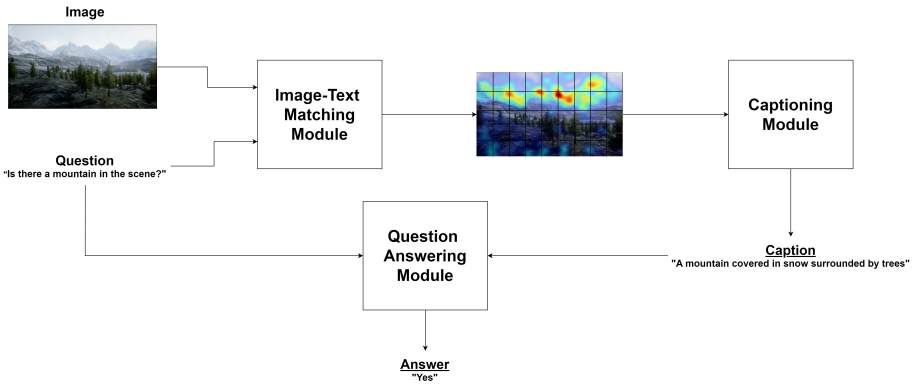

Fig. 3: The employed VQA architecture.

a specific function apiece, certain rules the GPT3.5 outputs must follow, the general goal of the task and tips on how to filter and extract information from captions. Additionally, to prevent hallucination [34], i.e., imaginative and fabricated outputs from the controller, we begin the prompt by informing LLM that it is in a game scenario, the commands serve as its controls and the goal is to provide a detailed description of the observed scene while looking out for any possible anomalies that could lead to hazardous situations. The list of commands is split into:

(i) Active perception commands
  (a) Move closer, to move 10 meters forward.
  (b) Move back, to move 5 meters backwards.
  (c) Move right, to move 10 meters to the right.
  (d) Move left, to move 10 meters to the left.
(ii) General control commands
  (a) Save position, to save the current position of the drone.
  (b) Ask a question, to ask exploratory questions.
  (c) I know enough, to return to the starting position.

Additionally, we divide the diverse list of rules the LLM must follow into:

(i) General Rules, to make sure LLM outputs the commands and questions correctly.
(ii) Active Perception rules, which ensure the proper movement of the drone.
(iii) Visual Question Answering rules, in order to utilize the captions and answers as efficiently as possible and optimize the procedure.

The propose pipeline consists of the following: an *active perception* mode, a *validation* mode and an *explanation* mode. Throughout active perception mode, the drone's camera takes snapshots of the observed scene and the controller asks questions while simultaneously issuing different movement commands. The process always starts with the question "What do you see?". Consequently, the VQA

model returns an answer, a caption and a percentage indicating if the caption matches the specific image to help distinguish between accurate and inaccurate captions. Through multiple diverse captions from different angles of the scene, the LLM model is able to gain knowledge and by leveraging its language understanding capabilities it is able to generalize and understand the context, as well as output possible safety measures for the specific scene. Then, during exploration mode, we encourage the LLM (by providing the appropriate prompt) to use the command *save position* whenever it deems it necessary in order to save the current drone position and revisit it during validation mode. The process continues until the LLM uses the command *I know enough* and transitions to validation mode.

During the validation mode, we ask the LLM to output a description and a caption of its current knowledge, along with which parts it wants to validate. We add random Gaussian noise to the saved positions, in order to gain different question-image pairs before inputting them to the VQA model again. In each new position, the controller asks one validating question for each targeted piece of information it wants to validate and we also save the question-image pairs which hold the highest matching score percentage for explanation mode. Afterwards, the controller compiles all the answers in each revisited position and leverages an ensemble approach to update the scene description and caption. In the end, the drone returns to its starting position outputting the final description, caption and the safety rules about the scene.

Finally, in order to provide the ability to explain the conclusions drawn by the developed pipeline, we extract the GradCAM's visualization from our VQA model in order to output attention maps on the validated images, as shown in Fig. 2. As a result, when the drone returns to its starting position it is able to output the question-image pairs through an attention mask, highlighting the parts of the image that lead to its decisions on the captioning and question-answering tasks.

4 Experimental Evaluation

All the experiments were conducted using the Airsim simulation environment [35]. It is built upon Unreal Engine 4 and consists of a physics engine and different environmental, vehicular and sensory models. By testing out the quadrotor vehicular model in multiple environments we can simulate a plethora of scenarios that provide physical and visual feedback adjacent to the real world. Specifically, our experiments take place in typical surveillance environments such as a mountain landscape, a lake, a public square and a snowy road, as shown in Fig. 4.

To quantitatively evaluate the performance of the proposed method we compute the caption-image matching score (using the VQA model) at the drone's spawn position and at every subsequent position revisited during the validation module. We then calculate the average caption-image matching score across all positions for ten independent experiments. The results are reported in Table 1,

Fig. 4: Four different environments were used for the conducted experiments: a mountain landscape, a snowy road, a public square and a lake.

where we compare the baseline score (directly using the description at the starting position of the drone), and the final validated result of the proposed method ("Proposed"). Our results indicate that in different environments, the proposed method consistently enhances the caption-image matching score, suggesting that the generated captions provide more relevant information that aligns well with the scene. Furthermore, we present the average run time required, to obtain a validated, detailed scene description with explainable attention maps. Given that the average experiment time is approximately 12 minutes and recognizing that such a duration is impractical in hazardous situations, we introduce a special rule in our prompt. This rule stipulates that whenever the proposed method detects a potential anomaly, it must immediately stop exploration and proceed with validation and result generation. By implementing this rule, we reduce the average experiment time to under 5 minutes in anomaly induced scenarios.

Table 1: Average image-caption matching score (calculated over ten runs) for each of the employed environments.

Environment	**Baseline**	**Proposed**	**Time of Experiment**
Mountain Landscape	0.384	**0.585**	**12mins 57secs**
Public Square	0.361	**0.699**	**12mins 28secs**
Snow road	0.458	**0.629**	**11mins 48secs**
Lake	0.451	**0.690**	**13mins 26secs**

Fig. 5: Example anomalies in the four different environments. Note that some anomalies are challenging to detect and require very careful inspection of the input frame.

Additionally, we assess our system's performance on the task of anomaly detection. By introducing potential hazards or dangerous elements, such as fires and car crashes, into each scene (refer to Fig. 5 for some example anomalies), we evaluate the baseline framework's ability to accurately identify anomalies, comparing it with the performance of our proposed system following the active perception and validation phases. We consider the system successful in anomaly detection when it identifies the anomaly in its captions in a coherent and grammatically logical manner. To evaluate the proposed method in scenes that contain anomalies, we deploy hazards in three distinct scenarios. Initially, we position a potential hazard within the range of the drone's spawn point. Subsequently, we increase the distance between the drone's spawn point and the hazard. Finally, we place the hazard in an obscured view from the initial drone spawn point necessitating movement to locate it. We conduct the experiments ten times for each environment and present the accuracy of anomaly detection (averaging the ten runs over the three setups), comparing the baseline and the proposed method, in Table 2.

These results indicate that the drone succeeded in providing a description and caption about the unknown scene whilst only relying on outputs from the VQA model in the form of text. Moreover, when hazardous anomalies are introduced, altering the scene to an unsafe condition, our system successfully identifies the danger and suggests necessary safety precautions. Finally, the proposed pipeline can also provide interpretable attention maps, leveraging GradCAM's capabili-

Table 2: Comparing anomaly detection accuracy between baseline and the proposed method.

Method	**Environment**	**Anomaly Detection Score**
Baseline	Mountain Landscape	0.53
Proposed	Mountain Landscape	**0.90**
Baseline	Public Square	0.43
Proposed	Public Square	**0.73**
Baseline	Lake	0.26
Proposed	Lake	**0.76**
Baseline	Snow	0.20
Proposed	Snow	**0.83**

ties, both for the intermediate and final questions/captions, which showcase the validated information in order for a human operator to assess. Two indicative examples are shown in Fig. 6, highlighting the improved explainability capabilities provided by the proposed method. Furthermore, in Table 3, we compare the captions provided by the baseline model with the captions provided by the proposed framework and in Table 4 we provide the detailed scene descriptions leveraged by our proposed framework. Note that in most cases the proposed method leads to a more accurate description. However, hallucinations can still occur despite the validation process. Increasing the number of examination points and/or adding additional validation steps could help further reduce these occurrences.

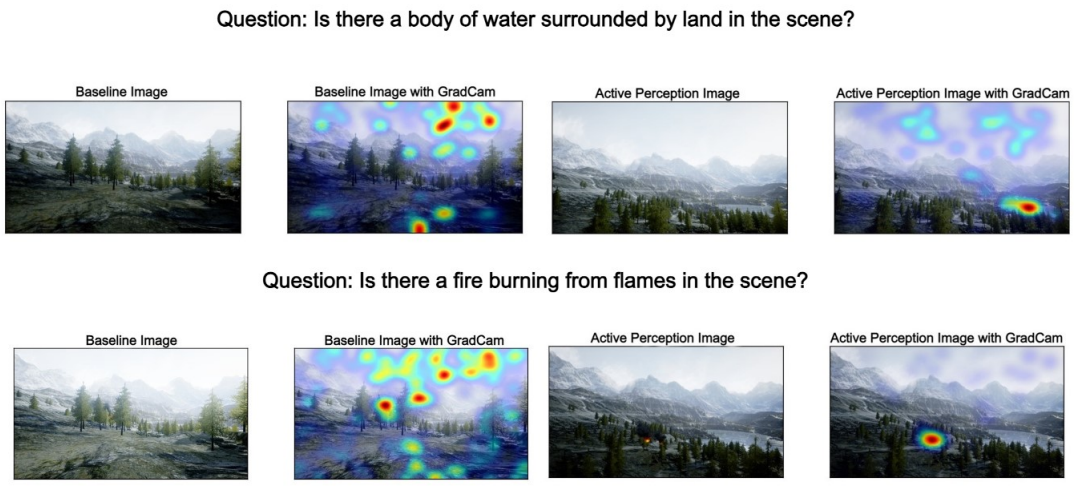

Fig. 6: Two examples for two different questions, indicating the additional explain ability capabilities that can be provided by the proposed pipeline.

Table 3: Caption examples provided by the baseline VQA model and the proposed model. We highlight the correct pieces of information with the color green, the wrong ones with the color red and the ambiguous ones with orange.

Scene	Baseline	Proposed
Mountain Landscape	A view of rocky mountain peaks that looks into the horizon	A serene mountainous landscape with mist, snow-capped mountains, and trees.
Snowy road in mountainside	The snowy mountain is covered in a thick blanket of snow.	A snowy mountain with a road leading into glacier water.
Public Square	A fountain park filled with lots of water.	A lively fountain park shrouded in dense fog with water shoots creating a mysterious atmosphere
Lake	A group of tall vegetation on a river.	A tranquil lake setting with ducks, tall vegetation, and lush green plants, offering a picturesque natural landscape.
Mountain Landscape with fire	A huge flame and a cloud of black smoke.	A devastating forest fire consumes the valley, threatening the green vegetation and trees in its path.
Lake with a car fire	The steam rises in the clouds on a foggy day.	A car crash has occured, with a truck damaged after crashing into a river emitting smoke, individuals trying to move the stuck truck.

Table 4: We showcase our methods ability to provide descriptions of the scenes, after the information was gathered through Active Perception and after it was validated through our validation module. We highlight the correct pieces of information with the color green, the wrong ones with the color red and the ambiguous ones with orange.

Environment	Proposed Final Description
Mountain Landscape	The aerial surveillance drone has captured a serene mountain landscape with trees covering its slopes. While there is no visible forest in the scene, a clear lake adds to the natural beauty of the surroundings. The absence of human activity enhances the peacefulness of the environment.
Mountain Landscape with fire	The aerial surveillance drone captures a dramatic scene with a group of mountains featuring rocky peaks in the background. In the foreground, a fire rages with red lava and flames, casting a fiery glow. On the left side, a majestic mountain stands tall, adding to the rugged landscape. Meanwhile, on the right side, another fire burns with smoke billowing into the sky. The background displays a computer artwork, adding a surreal touch to the overall view.
Snowy road in mountainside	The scene depicts a tranquil snowy landscape with no specific objects or anomalies present. The serene setting is characterized by the peacefulness of the snow-covered terrain and the absence of any notable features.
Snowy road in mountainside with car crash	The scene depicts a snowy road with a truck traveling on it. The road is covered in snow, and there is a mountain nearby covered in heavy snow. The presence of the truck on the snowy road indicates a potential hazardous situation that needs to be approached with caution.
Public Square	The scene features a round, red tiled courtyard enveloped in fog, creating an eerie and mysterious atmosphere. The fog obscures the surroundings, adding to the sense of obscurity and intrigue. The digital object, previously mentioned, is no longer present in the scene leaving behind a solitary and enigmatic courtyard.
Public Square with fire	The scene features a small fountain with water spraying, and an outdoor fountain with a fire display, and a fire torch made of metal. Both the small fountain and fire display have been confirmed to be present in the scene. The fire torch made of metal is also part of the scene, adding to the overall ambiance.
Lake	The scene portrays a tranquil river flowing with ripples at its center. Along the riverbank, the trees stand tall and healthy, framing the water's edge without any nearby structures interrupting the natural beauty. Across the river lies a park merging into a dense forest, enhancing the scene's idyllic charm. A blanket of fog envelops the surroundings, lending an air of mystery and serenity to the landscape.
Lake with fire	The scene features a body of water with a small boat floating in the middle. In front of the boat, a tree is engulfed in flames, emitting orange burning flames. The fire has spread to the bush tucker on a field with trees. However, there is no floating island engulfed by flames as previously mentioned. Smoke rises from the burning objects, creating a hazardous environment.

5 Conclusion

In this paper, we presented a novel framework that employs LLMs to actively collect information and detect anomalies, even in unprecedented situations. We propose a method where two deep learning models engage in dialogue to control a drone and improve anomaly detection accuracy. We test our approach in a

realistic simulation environment, where the LLM follows natural language commands to move the drone, while a VQA model answers questions about images. By combining these models, the LLM asks questions while guiding the drone through the scene, providing a unique way to improve perception accuracy, as well as detect potential anomalies. At the same time, by leveraging the explainability capabilities of the employed VQA model, the proposed method can also further improve the explainability of the perception process. By providing four different types of scenes, with different hazardous situations in them and without requiring any fine-tuning or retraining of the models, we demonstrate the potential of the proposed method for handling open-ended adaptation in-the-wild. Additionally, to the best of our knowledge, there is currently no other established way to implement and evaluate active perception in unstructured open-world setups. Therefore, this work opens several research directions, including effective evaluation of approaches that extend beyond static perception and pave the way for applications in other areas as well.

Acknowledgements The work presented here has been partially supported by the RoboSAPIENS project funded by the European Commission's Horizon Europe programme under grant agreement number 101133807. This publication reflects the authors' views only. The European Commission is not responsible for any use that may be made of the information it contains.

References

1. B. Mishra, D. Garg, P. Narang, and V. Mishra, "Drone-surveillance for search and rescue in natural disaster," *Computer Communications*, vol. 156, pp. 1–10, 2020.
2. A. Chriki, H. Touati, H. Snoussi, and F. Kamoun, "Uav-based surveillance system: an anomaly detection approach," in *Proceedings of the IEEE Symposium on Computers and Communications (ISCC)*, 2020, pp. 1–6.
3. R. Gasparini, A. D'Eusanio, G. Borghi, S. Pini, G. Scaglione, S. Calderara, E. Fedeli, and R. Cucchiara, "Anomaly detection, localization and classification for railway inspection," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2021, pp. 3419–3426.
4. X. Zhai, K. Liu, W. Nash, and D. Castineira, "Smart Autopilot Drone System for Surface Surveillance and Anomaly Detection via Customizable Deep Neural Network," ser. IPTC International Petroleum Technology Conference, vol. Day 2 Tue, January 14, 2020, 01 2020, p. D021S053R001.
5. E. Unlu, E. Zenou, N. Riviere, and P.-E. Dupouy, "An Autonomous Drone Surveillance and Tracking Architecture," in *2019 Autonomous Vehicles and Machines Conference, AVM 2019*, vol. 2019, Jan. 2019, pp. 35–1–35–7.
6. R. Bajcsy, Y. Aloimonos, and J. K. Tsotsos, "Revisiting active perception," *Autonomous Robots*, vol. 42, pp. 177–196, 2018.
7. N. Saito, T. Ogata, S. Funabashi, H. Mori, and S. Sugano, "How to select and use tools?: Active perception of target objects using multimodal deep learning," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2517–2524, 2021.
8. T. Manousis, N. Passalis, and A. Tefas, "Enabling high-resolution pose estimation in real time using active perception," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2023, pp. 2425–2429.

9. A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra, and D. Parikh, "Vqa: Visual question answering," *arXiv:1505.00468*, 2016.
10. Y. Deng*, D. Guo*, X. Guo, N. Zhang, H. Liu, and F. Sun, "Mqa: Answering the question via robotic manipulation," in *Robotics: Science and Systems XVII*, ser. RSS2021. Robotics: Science and Systems Foundation, 2021.
11. D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi, "Iqa: Visual question answering in interactive environments," *arXiv:1712.03316*, 2018.
12. A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Embodied question answering," *arXiv:1711.11543*, 2017.
13. S. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor, "Chatgpt for robotics: Design principles and model abilities," *arXiv:2306.17582*, 2023.
14. M. Lamine, . Tazir, M. Mancas, and T. Dutoit, "From words to flight: Integrating openai chatgpt with px4/gazebo for natural language-based drone control," *Proceedings of the 13th International Workshop on Computer Science and Engineering*, 2023.
15. Y. Ye, H. You, and J. Du, "Improved trust in human-robot collaboration with chatgpt," *arXiv:2304.12529*, 2023.
16. J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," *arXiv:2209.07753*, 2023.
17. C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan, "Visual chatgpt: Talking, drawing and editing with visual foundation models," *arXiv:2303.04671*, 2023.
18. Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face," *arXiv:2303.17580*, 2023.
19. S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, "Next-gpt: Any-to-any multimodal llm," *arXiv:2309.05519*, 2023.
20. M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," *arXiv:2109.12098*, 2021.
21. A. Bucker, L. Figueredo, S. Haddadin, A. Kapoor, S. Ma, and R. Bonatti, "Reshaping robot trajectories using natural language commands: A study of multi-modal data alignment using transformers," *arXiv:2203.13411*, 2022.
22. S. Stepputtis, J. Campbell, M. Phielipp, S. Lee, C. Baral, and H. B. Amor, "Language-conditioned imitation learning for robot manipulation tasks," *arXiv:2010.12083*, 2020.
23. D. Zhu, J. Chen, K. Haydarov, X. Shen, W. Zhang, and M. Elhoseiny, "Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions," *arXiv:2303.06594*, 2023.
24. N. Rotstein, D. Bensaid, S. Brody, R. Ganz, and R. Kimmel, "Fusecap: Leveraging large language models for enriched fused image captions," *arXiv:2305.17718*, 2023.
25. M. Levy, R. Ben-Ari, N. Darshan, and D. Lischinski, "Chatting makes perfect: Chat-based image retrieval," *arXiv:2305.20062*, 2023.
26. R. Ricci, Y. Bazi, and F. Melgani, "Machine-to-machine visual dialoguing with chatgpt for enriched textual image description," *Remote Sensing*, vol. 16, no. 3, 2024.
27. Y. Hu, H. Hua, Z. Yang, W. Shi, N. A. Smith, and J. Luo, "Promptcap: Prompt-guided task-aware image captioning," *arXiv:2211.09699*, 2023.
28. Z. Yu, X. Ouyang, Z. Shao, M. Wang, and J. Yu, "Prophet: Prompting large language models with complementary answer heuristics for knowledge-based visual question answering," *arXiv:2303.01903*, 2023.
29. S. Ravi, A. Chinchure, L. Sigal, R. Liao, and V. Shwartz, "Vlc-bert: Visual question answering with contextualized commonsense knowledge," *arXiv:2210.13626*, 2022.

30. A. M. H. Tiong, J. Li, B. Li, S. Savarese, and S. C. H. Hoi, "Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training," *arXiv:2210.08773*, 2023.
31. J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," *arXiv:2201.12086*, 2022.
32. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.
33. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Proceedings of the Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.
34. Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, L. Wang, A. T. Luu, W. Bi, F. Shi, and S. Shi, "Siren's song in the ai ocean: A survey on hallucination in large language models," *arXiv:2309.01219*, 2023.
35. S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," *arXiv:1705.05065*, 2017.