remote sensing

MDPI

Article

Machine-to-Machine Visual Dialoguing with ChatGPT for Enriched Textual Image Description

Riccardo Ricci [1,†], Yakoub Bazi [2,†] and Farid Melgani [1,*]

[1] Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy; riccardo.ricci-1@unitn.it
[2] Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 4545, Saudi Arabia; ybazi@ksu.edu.sa
* Correspondence: melgani@disi.unitn.it
† These authors contributed equally to this work.

Abstract: Image captioning is a technique that enables the automatic extraction of natural language descriptions about the contents of an image. On the one hand, information in the form of natural language can enhance accessibility by reducing the expertise required to process, analyze, and exploit remote sensing images, while on the other, it provides a direct and general form of communication. However, image captioning is usually restricted to a single sentence, which barely describes the rich semantic information that typically characterizes remote sensing (RS) images. In this paper, we aim to move one step forward by proposing a captioning system that, mimicking human behavior, adopts dialogue as a tool to explore and dig for information, leading to more detailed and comprehensive descriptions of RS scenes. The system relies on a questions–answers scheme fed by a query image and summarizes the dialogue content with ChatGPT. Experiments carried out on two benchmark remote sensing datasets confirm the potential of such an approach in the context of semantic information mining. Strengths and weaknesses are highlighted and discussed, as well as some possible future developments.

Keywords: ChatGPT; image captioning; visual question answering (VQA); visual question generation (VQG); visual dialoguing

check for updates

Citation: Ricci, R.; Bazi, Y.; Melgani, F. Machine-to-Machine Visual Dialoguing with ChatGPT for Enriched Textual Image Description. *Remote Sens.* **2024**, *16*, 441. https://doi.org/10.3390/rs16030441

Academic Editors: Chiman Kwan, Moulay A. Akhloufi and Mozhdeh Shahbazi

Received: 19 September 2023
Revised: 28 December 2023
Accepted: 19 January 2024
Published: 23 January 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Remote sensing images are characterized by rich and detailed semantics, which makes them difficult and time-consuming to analyze, especially for non-experts. Traditional approaches of automatic information extraction such as scene classification, semantic segmentation, and others can extract very low-level information that, usually, must be refined or reasoned further to translate it into a more useful source of information for the final user. In this context, remote sensing (RS) image captioning aims at bridging the need for a more complete and straightforward source of information by extracting information directly in the form of natural language, which is more interpretable by a human user. Image captioning teaches machines to maximize the probability of generating the ground truth description through end-to-end training of large neural networks. This can be difficult to achieve due to the enormous space of possibilities that image–description couples represent. Furthermore, image captioning often extracts a single sentence, which may be not enough to capture the rich and detailed semantics of (especially) RS images. In this paper, we try to model the image captioning process by proposing a different paradigm that, instead of directly optimizing a set of weights conditioned on a set of supervised training pairs, tackles the description problem in a more human-like way, namely through dialoguing. We argue that dialogue is beneficial in two ways: (1) it decomposes the problem into subproblems represented by various questions, whose solutions can be easier; (2) it enables exploring to a deeper degree the semantic content hidden in an image; and (3) it is less sensitive to the common problem of error accumulation in image captioning [1] in which an error in the prediction of a word will most likely influence all the other words by breaking down the description and so reduces the errors to sub-problems. In particular, we explore different solutions for enriched remote sensing image description under the paradigm of "description through dialogue", providing strengths and weaknesses of each, as well as directions for further research. The paper is organized as follows: in Section 2, after an introduction to the fundamentals of image captioning (IC) in the context of RS images, and various related research areas such as visual question answering (VQA) and visual question generation (VQG), we introduce our proposed paradigm of description through dialogue by bridging these different research areas. Specifically, we outline three different variants of the dialoguing system and the metrics used for assessment. Section 3 reports the results obtained by the various approaches, while Section 4 reports quantitative and qualitative results for all three solutions of the dialoguing system on both datasets, followed by a general discussion about the performance in different contexts. Finally, Section 5 outlines open issues and future directions, concluding the paper.

1.1. Related Research Areas

First, we want to introduce various techniques for the automatic extraction of information from images using natural language. Starting with the most seminal area of generative-based image captioning, additional research areas, namely visual question answering (VQA) and visual question generation (VQG), are also introduced. It is noteworthy that, in the following subsections, the concept of token refers to a generic piece of a text sequence, which can be different depending on the tokenization scheme.

1.1.1. RS Image Captioning (RSIC)

Remote sensing image captioning (RSIC) is the task of automatically describing a remote sensing image using natural language. RSIC is being applied to tackle multimodal image retrieval and change captioning [2] and can be used in other areas such as traffic command or rescue scenarios [3] by providing automatic text or voice descriptions of unmanned aerial vehicle (UAV) images. Generative-based remote sensing image captioning witnessed a similar evolution compared to the natural image counterpart. RSIC commonly relies on an encoder-decoder architecture, in which the encoder is devoted to the extraction of discriminative features from the image, and the decoder, conditioned by the extracted features, is employed to generate a description. For a practical example, the seminal paper [4] introduces an encoder-decoder architecture employing a pre-trained convolutional neural network (CNN) to extract image features and an auto-regressive decoder (RNN) to generate the description word-by-word conditioned on both the image features and the previously predicted words. Soon, building on this first approach, advances in encoder (deeper architectures [5]), decoder (gated recurrent unit (GRU [6], SVM-based decoder [7]), and in the interaction between the text and image features (multilevel features [8], attention mechanism [9–11]), enabled researchers to push further the limits of RSIC performance. Recently, the introduction of transformer architecture led to fully attention-based captioning systems [12]. All image captioning approaches based on deep neural networks model their task as an auto-regressive process, in which the next token is generated conditioned on the image features and the previously predicted tokens. More specifically, given an image–description pair, traditional image captioning algorithms try to maximize the following probability distribution:

$$P(\mathbf{T}|I) = \prod_{k=1}^{L} P(t_k|\mathbf{T}_{1\ldots(k-1)}, I) \tag{1}$$

where $\mathbf{T}$ is a ground-truth description composed of L tokens and I is the vector of image features. Given a collection of image–description pairs, the network is enforced to increase the probability of assigning the correct description to every image in the dataset. Afterward, the trained network associates an input image with the highest fitting sequence of words according to the learned distribution. Such a paradigm has been widely applied in RSIC but presents some limitations. The first is that all datasets for RSIC are composed of images and short descriptions, which barely is enough to describe the wide variety of objects and concepts that characterize RS images. The second is the difficulty in learning such a complex distribution with the relatively small RS datasets that exist in the literature [4,8,13]. The cost associated with the collection and labeling of data for remote sensing image captioning makes it unfeasible to deliver comprehensive and large datasets adequate for an algorithm to reliably learn a good approximation of the true distribution, even if some efforts are being made in this direction [13]. This makes algorithms for remote sensing image captioning less robust and general, thus restricting their applicability.

1.1.2. RS Visual Question Answering (RSVQA)

The branch of visual question-answering research started shortly after image captioning [14]. The task in VQA is to answer user-defined questions by looking at the contents of an image. This enables users to interact with VQA systems to analyze images, retrieving specific information using natural language questions. In the context of remote sensing, a VQA system can be compared to an expert who can provide answers on specific concepts about remote sensing images. VQA methods may be subdivided into two main categories: closed-ended and open-ended answering. Closed-ended answering is defined as a classification task on a closed set of possible answers, modeling a probability distribution as:

$$P(a|\mathbf{Q}, I) \tag{2}$$

where a is one answer between a set of predetermined fixed answers, $\mathbf{Q}$ is the sequence of tokens representing the natural language question, and I is the vector of image features. In open-ended visual question answering, similarly to the task of image captioning, the answer is a sequence of tokens generated in an auto-regressive manner. The probability distribution modeled by open-ended VQA algorithms is reported in Equation (3), which closely resembles that modeled by a captioning system, except for the further conditioning given by the question Q:

$$P(\mathbf{A}|\mathbf{Q}, I) = \prod_{k=1}^{L} P(A_k|A_{1..k-1}, Q, I) \tag{3}$$

RS visual question answering (RSVQA) has been tackled for the first time in [15] using a closed-form answering paradigm. The authors used OpenStreetMap and a template-based question generation process to gather question–answer pairs for RS images, creating the first dataset for remote sensing visual question answering [15]. However, as shown by them, RSVQA suffers from the complexity of remote sensing scenes, for example, when dealing with the counting and numbering of objects. Despite being very recent, the field of RSVQA has witnessed some advancements, from the introduction of the attention mechanism [16], to ways that allow the algorithm to account for subsequent questions [17], to the leveraging of large language models in the context of prompt answering [18] and the adoption of the transformer architecture for both images and text and their fusion [19].

1.1.3. RS Visual Question Generation (RSVQG)

Visual question generation is a novel research area, which aims at automating the questioning process by generating coherent and sound questions for an input image. Visual question generation can be useful, for example, to gather data for visual question answering, where data labeling is very expensive and time-consuming. Other areas where visual question generation can be useful are in the field of child education, interactive lectures, and visual conversations [20]. VQG methods can be divided into three groups, namely rule-based, template-based, and generative-based visual question generation. In rule-based VQG, some specific syntactical rules are used to build questions from image descriptions [21]. Template-based VQG proposes to generate questions using templates and a set of objects, attributes, and relations. Those can be fixed [22] or automatically recognized in the input image using some visual detectors [23]. Generative-based VQG shares similarities with image captioning, in which both are required to generate text in an open-handed way. One peculiarity of generative-based VQG that differentiates it from image captioning is the requirement to generate not just one, but a set of plausible and coherent sequences of text (questions) for the same input image. Different generative-based methods have been envisioned to generate different questions for the same image. For example, in [22], questions are generated that explicitly explore objects detected in the image, further constraining the generation to predefined types of questions. In [24], different questions are created, starting from regional captions and adding the conditioning given by the question type. In [25], multiple question generation is addressed in an end-to-end leveraging the variational autoencoder framework. Despite advances in generative-based visual question generation, the application to remote sensing images is still an open area of research. From what we know, ref. [26] is the only work in the literature exploring generative-based visual question generation for RS images. In [26] the authors propose a transformer-based architecture to generate different plausible questions for RS images. They use questions from the dataset proposed in [15] to train the model in a fully supervised fashion.

2. Materials and Methods

In this work, we propose a system that can extract information from images, condensing it in a descriptive paragraph of natural language. The system building blocks are derived from the previously introduced research areas, bridging visual question generation with visual question answering. Its general idea, depicted in Figure 1, is to establish a machine-to-machine (M2M) visual dialoguing (VD) that requires no (or little) concrete human intervention and that enables information digging from images. The collected information is summarized at the end of the dialogue to generate the final output. This way of approaching the problem is different from standard image captioning. In image captioning, there is no real exploration of the information in an adaptive way, but all the information is already condensed in a single round output description, which can reduce the extraction of information. Under the dialoguing paradigm, the exploration is performed in steps; each question–answer pair allows the machine to explore more of the image content, possibly building on previously extracted knowledge to decide which follow-up questions to ask to maximize the collection of information. To formalize the idea, we envision a system fed by an input image and composed of two machines, one that is devoted to the task of asking questions, and the other that, based on a given question, provides an appropriate answer. This conceptual division is useful to obtain the idea of the actors involved in the process since it is not necessarily required that those be two distinct machines. Indeed, different variants can emerge, for example, with a single machine that provides both questions and answers or with predefined template questions. In this paper, we will explore different preliminary solutions, but as we expect, more can emerge in the future. We will explore some strengths and drawbacks of different strategies, trying to set the path for more exploration of this promising but challenging paradigm.

We present three different solutions to the dialoguing system. The first, called “Open-ended dialogue”, is based entirely on the proposal outlined in [27]. We adopt the same configuration, experimenting with zero-shot dialoguing for remote-sensing images. The second, called “Closed-form dialogue”, uses a predefined set of questions and generates a fixed dialogue for each image. The third, called “Closed-form dialogue with context”, inherits the dialoguing approach of the previous closed-form solution but proposes a way to integrate the context given by the previous questions and answers in the answering process.

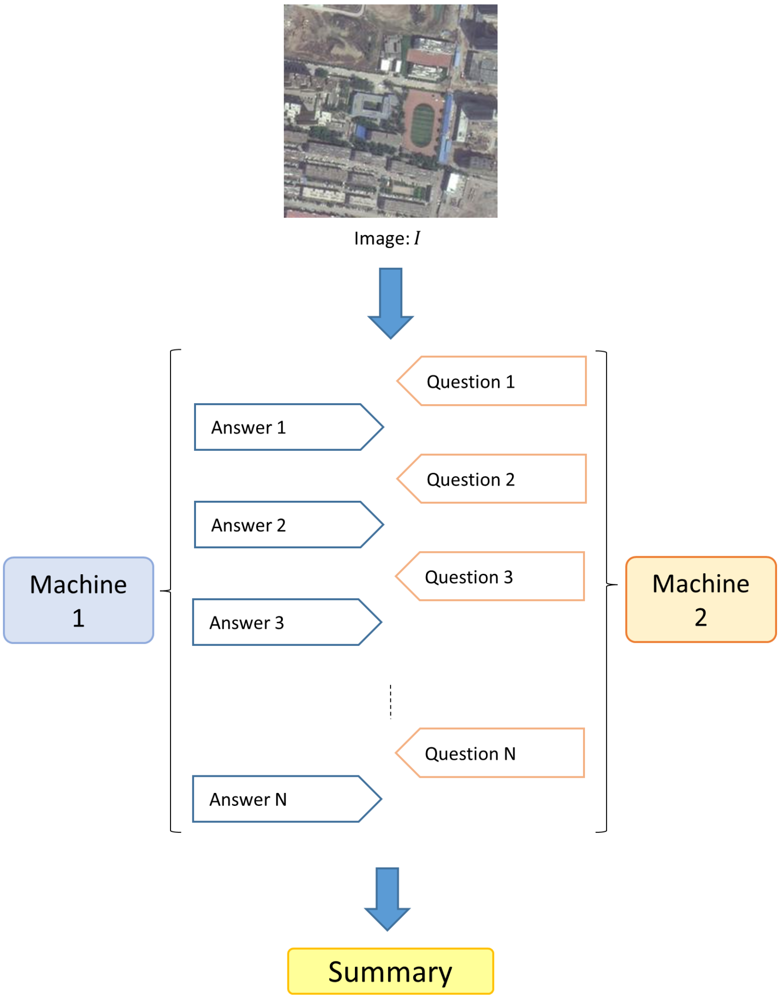


Figure 1. Conceptual representation of the machine-to-machine visual dialoguing (M2M-VD) paradigm.

2.1. Open-Ended Dialogue

In open-ended dialogue, the idea is to have two separate machines, one that produces questions and the other that answers. In this work, we adopt the method proposed in [27], in which Blip-2 [28] is used to answer questions generated by ChatGPT [29]. Blip-2 is a multimodal architecture that bridges the visual and text modalities by connecting a frozen image encoder to a frozen large language model (LLM). The goal is to preserve the reasoning ability of large language models while injecting image understanding by adaptively merging the two modalities through a so-called "query transformer". The alignment between the image encoder and the LLM follows a two-step procedure, using a corpus of 129M of images with corresponding text descriptions. In the first step, alignment of the extracted visual features is performed by connecting the query transformer to the frozen image encoder.

In detail, the authors make use of three pre-training objectives, namely (1) image-text contrastive learning (ITC), (2) image-grounded text generation (ITG), and (3) image-text matching (ITM). ITC aims at enhancing the mutual information between image–text representations of corresponding (positive) pairs while reducing that of non-corresponding (negative) pairs. ITG trains the query transformer to generate texts conditioned on the input image. This objective forces the queries to capture the most meaningful information from the frozen image encoder representation to generate the text. ITM consists of a binary classification of whether the image corresponds to the text (positive pair) or not (negative pair). The prediction is achieved by generating logits' scores from each query vector in the

Query Transformer and averaging to obtain the overall matching result. In the second step, the authors connect the query transformer to a frozen large language model. The goal of this stage is to align the large language model in generating text conditioned on the queries extracted by the query transformer. According to the authors, keeping the models frozen during alignment can mitigate the catastrophic forgetting problem, therefore preserving the ability to perform prompt-based text generation while including conditioning information from the image.

ChatGPT [29] is a fine-tuned version of GPT-3.5, and it belongs to the family of large language models. GPT-3.5 has been pre-trained to perform next token prediction in a self-supervised fashion on a huge amount of textual data collected from the web. Starting from the pre-trained GPT-3.5 checkpoint, in [30], authors use reinforcement learning from human feedback (RLHF) to enforce the instruction following capability. This passage equipped ChatGPT with superior capability in language understanding, generation, interaction, and reasoning.

Figure 2 depicts the block scheme for the open-ended dialogue system. In this framework, the question generation process is dynamically controlled by ChatGPT relying on the previous dialogue history (questions and answers). Specifically, the following steps are performed:

1. A fixed prompt "Describe this image in detail" is used to spark the conversation.
2. Blip-2 answers with a description of the image.
3. ChatGPT generates a question based on the description.
4. Blip-2 answers by looking at the image.
5. ChatGPT, using the context of the first description and the dialoguing history, produces another question to further explore the image contents.
6. Blip-2 answers by looking at the image.

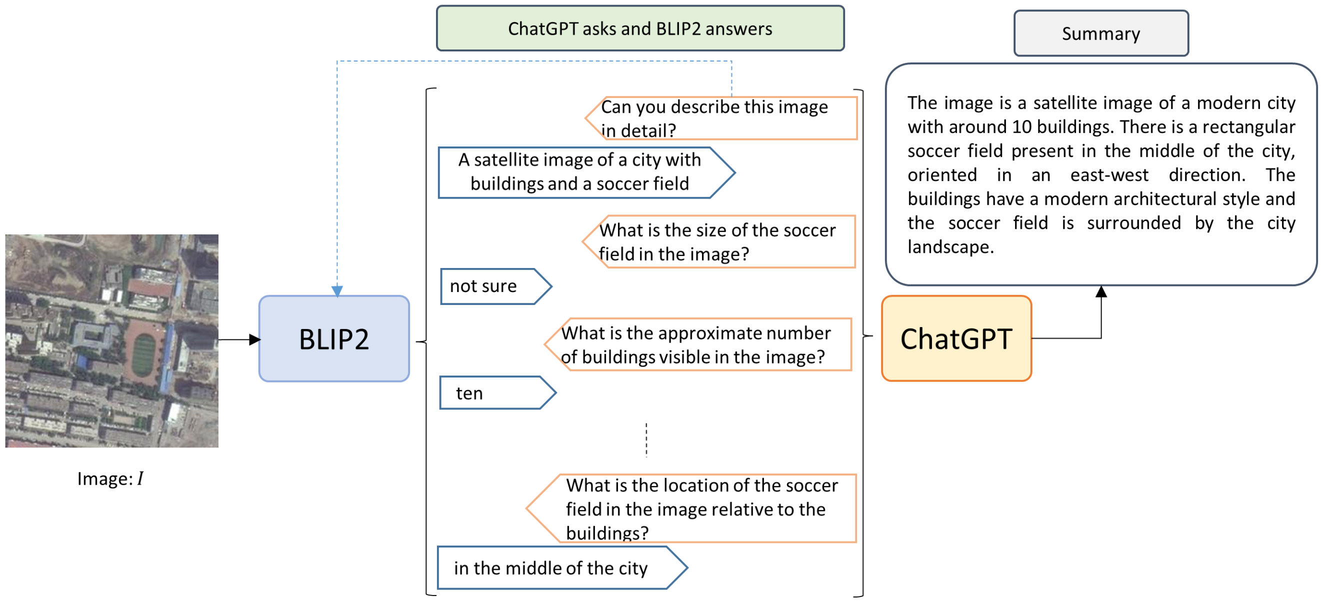
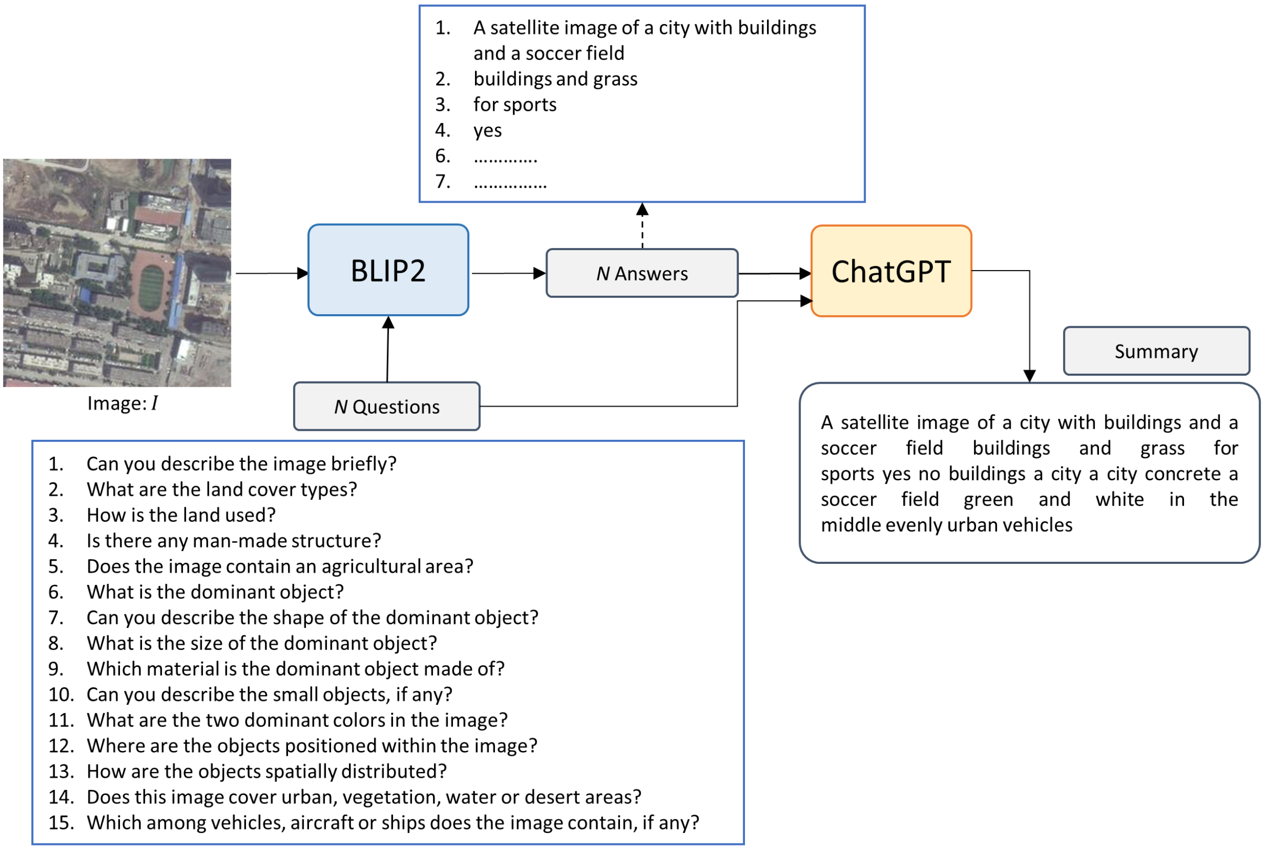

ChatGPT asks and BLIP2 answers
Summary
Can you describe this image in detail?
A satellite image of a city with buildings and a soccer field
What is the size of the soccer field in the image?
not sure
What is the approximate number of buildings visible in the image?
ten
What is the location of the soccer field in the image relative to the buildings?
in the middle of the city
The image is a satellite image of a modern city with around 10 buildings. There is a rectangular soccer field present in the middle of the city, oriented in an east-west direction. The buildings have a modern architectural style and the soccer field is surrounded by the city landscape.
BLIP2
ChatGPT
Image: I

Figure 2. Open-ended dialogue block scheme.

Points (5) and (6) are repeated N times to simulate a dialogue between the questioning machine and the answering machine. In our experiments, we simulated a total of 10 rounds of questions and answers, so $N = 9$. At the end of the dialogue, ChatGPT is also leveraged to summarize the content of the conversation and generate the final description.

2.2. Closed-Form Dialogue

In closed-form dialogue, we propose an alternative to alleviate some of the drawbacks of the open-ended pipeline. The major weakness of our strategy is the generation of follow-up questions, which heavily relies on the first description generated by Blip-2. Therefore, if this is incorrect, the dialogue may evolve completely incorrectly. One way to alleviate this problem is to decouple the question generation from the dialogue history. In closed-form dialogue, a predetermined set of questions is used for every image. Closed-form dialoguing can be useful in situations in which (1) the user needs to collect specific information regarding the scenes and is thus able to define a fixed set of targeted questions or (2) the first description by Blip-2 does not capture enough/correct information about the image to process meaningful subsequent questions. In the second case, the predefined set of questions, if well-designed for the target images, allows for a more consistent extraction of information and thus for more reliable image descriptions. As depicted in Figure 3, in closed-form dialogue, Blip-2 is used to answer the questions listed in the fixed set, one by one. Then, the set of questions and corresponding answers composing the dialoguing history is summarized by ChatGPT to produce a descriptive paragraph.



1. A satellite image of a city with buildings and a soccer field
2. buildings and grass
3. for sports
4. yes
6. …………
7. ……………

A satellite image of a city with buildings and a soccer field buildings and grass for sports yes no buildings a city a city concrete a soccer field green and white in the middle evenly urban vehicles

1. Can you describe the image briefly?
2. What are the land cover types?
3. How is the land used?
4. Is there any man-made structure?
5. Does the image contain an agricultural area?
6. What is the dominant object?
7. Can you describe the shape of the dominant object?
8. What is the size of the dominant object?
9. Which material is the dominant object made of?
10. Can you describe the small objects, if any?
11. What are the two dominant colors in the image?
12. Where are the objects positioned within the image?
13. How are the objects spatially distributed?
14. Does this image cover urban, vegetation, water or desert areas?
15. Which among vehicles, aircraft or ships does the image contain, if any?

Figure 3. Closed-form dialogue block scheme.

2.3. Closed-Form Dialogue with Context

In closed-form dialogue with context, we try to move one step forward in exploring the interaction between the image and the previous dialogue, and how it can influence the generation of answers for subsequent questions. In Figure 4, we propose an encoder-decoder architecture based on Clip [31] and GPT-2 [32] models. We rely on a ViT-base-16 to encode the image I into a sequence of tokens $f(I)$. Then, we concatenate these tokens with the first question $[Q_1]$ and feed them as input to an auto-regressive distil-GPT-2 [33] decoder to predict the answer $\hat{A}_1$ in an auto-regressive manner. To form the contextual information for the subsequent question Q_2, we concatenate the previous question and its predicted answer as follows $[f(I), Q_1, \hat{A}_1, Q_2]$ to generate the answer $\hat{A}_2$. The process continues until the generation of the last answer $\hat{A}_N$ for the question Q_N with the following contextual information $[f(I), Q_1, \hat{A}_1, Q_2, \ldots, Q_{N-1}, \hat{A}_{N-1}, Q_N]$. After that, we feed the concatenation of all questions and answers to ChatGPT to generate the summary as in the previous methods. It is worth noting that fine-tuning the weights of this encoder-decoder architecture (see Figure 5) is achieved by optimizing the so-called cross-entropy loss on the training set, which can be given for one dialogue related to the image I as follows:

$$L = \frac{1}{N}\sum_{k=1}^{N} Cross\_entropy(A_k, \hat{A}_k) \quad (4)$$

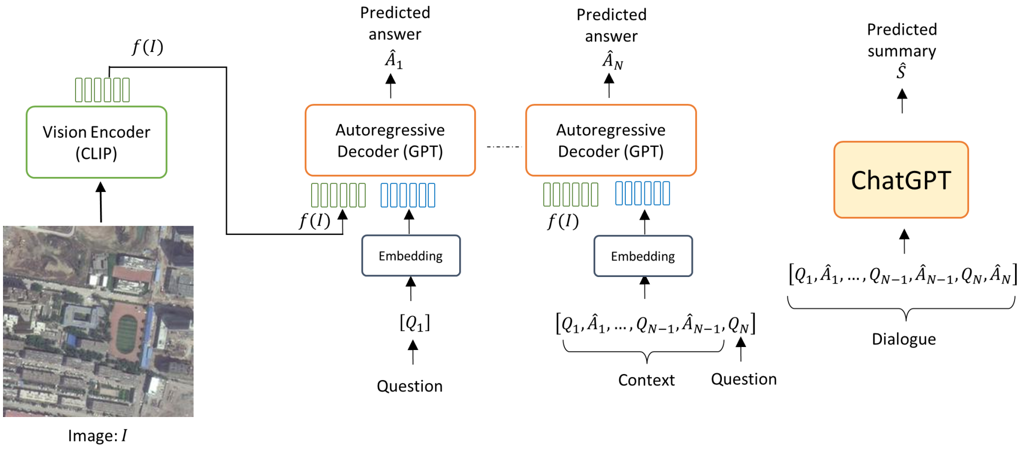


Figure 4. Block scheme of closed-form dialogue with context during inference phase.

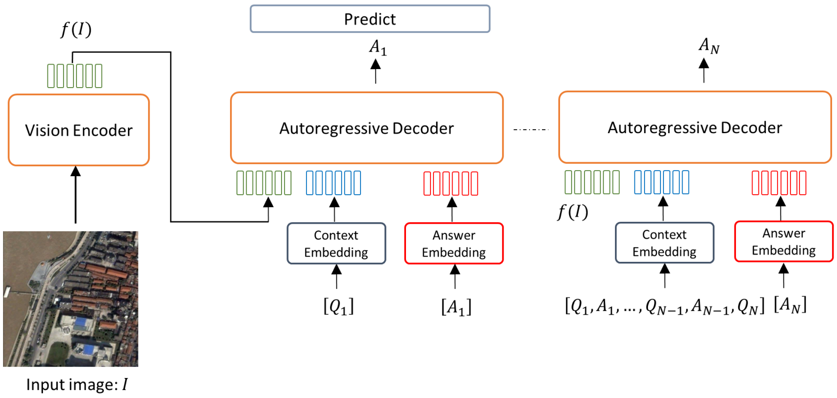


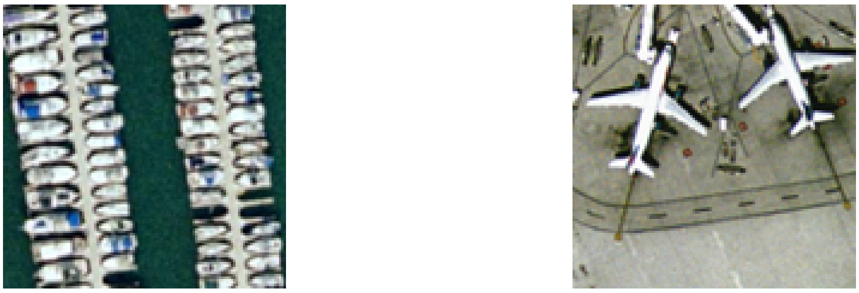
Figure 5. Closed-form dialogue with context during training phase.

2.4. Dataset and Metrics

Experiments have been carried out on two popular datasets for remote sensing image captioning, namely the RSICD dataset [8] and UCM-Captions dataset [4]. The RSICD dataset contains 10921 images of size 224 × 224 with varying spatial resolutions. Each image is annotated with a set of five captions. For some images, replication has been used to artificially augment the number of captions up to five. UCM-Captions includes 2100 aerial images of size 256 × 256 with a spatial resolution of one foot. This dataset is defined based on the UC Merced Land Use dataset, in which each image is associated with 1 of 21 land-use classes. Each image in the UCM-Captions dataset was annotated with five captions as well. Although five captions per image are available, in both datasets, captions of images belonging to the same class are very similar. Some examples of image–text couples for both datasets can be found in Figure 6. As stated in [27,34], assessing the performance in dialogue-based image description is a daunting problem, mainly because of the lack of ground-truth summaries to compare with. As explored in [27], the use of traditional metrics leads to huge drops in performance. In both these studies, authors use human assessment to evaluate the descriptions and the dialogues, which is expensive and time-consuming.

- This square playground is made up of two tennis courts coloured with green.
- Two small courts locate in a large grass.
- Two tennis courts surrounded by green meadows are near to several buildings.
- Two tennis fields are surrounded by green meadows.
- Two tennis courts are surrounded by a large piece of green meadow.

- A viaduct divided into some beige ring area and two gray straight viaduct.
- Viaduct is located in dark green vegetation area.
- A viaduct through the dark green plants and dark brown land.
- A pool with black water inside while surrounded by many dark green plants.
- Many green trees are near a viaduct.



- Lots of boats docked in lines at the harbor.
- Lots of boats docked neatly at the harbor.
- Many boats docked neatly at the harbor and the water is deep blue.
- Many boats docked neatly at the harbor and just a few positions are free.
- Lots of boats docked in lines at the harbor and just a few positions are free.

- There are two airplanes at the airport.
- Two airplanes are stopped at the airport.
- Two white airplanes are stopped at the airport.
- Two white airplanes are stopped at the airport with some cars beside it.
- Two white airplanes are stopped at the airport with some cars beside it.

Figure 6. Examples of images and corresponding captions from the RSICD and UCM-Captions datasets.

Therefore, we choose to experiment with other types of metrics, addressing the strengths and weaknesses of each in evaluating such results. The first is a reference-free metric called CLIPScore [35], another is text-to-image retrieval, and the last is a novel proposal, based on the generation of synthetic images from the summaries and comparison with the ground-truth images. CLIPScore is based on Clip [31], a multimodal deep neural network that learns shared vector representations of text and images. Clip has been trained on 400 M image–text pairs to align the representations of corresponding pairs while misaligning representations of mismatched pairs. In the contrastive learning paradigm used by Clip, "align" means to increase the cosine similarity between image and text representations. Leveraging the pre-trained Clip network, CLIPScore measures image–text compatibility using the following formula:

$$ClipScore(\vec{c},\vec{v}) = 2.5[max(cosine(\vec{c},\vec{v}),0)] \tag{5}$$

where $\vec{c}$ is an n-dimensional vector representation of the text and $\vec{v}$ an n-dimensional vector representation of the image. In the experiments, we adopted the Clip-ViT-L/14 variant, obtaining 512-dimensional vectors. The coefficient of 2.5 is used to stretch the measure in the range $[0-1]$. Higher scores have been shown to correlate with better descriptions of images. To obtain a measure for an entire dataset, the simple average for all samples is used.

For the second metric, we leverage text-to-image retrieval as a prompt of description goodness. The motivation for this choice is that if with a generated text one can isolate the image from which it has been created from the other images in the dataset, this means that the text must be more descriptive and coherent with the image itself. Indeed, suppose that the same description is generated for every image, then the retrieval is purely a matter of chance. The more the descriptions are targeted to the relative image, the more the description can isolate an image from the rest. We choose to experiment with text-to-image retrieval using Clip embeddings. Specifically, we experiment with Clip-ViT-L/14. We will report three recall metrics, R@1, R@5, and R@10, in decreasing order of strictness. More in detail, the recall measure is computed as follows. We have a dataset of N images, and we generate a description for image i. Using Clip-ViT-L/14, we encode all the images in the dataset and the generated description, yielding I and $\vec{t}$. I is a matrix of size $N \times 768$, where $\vec{I_j}$ is the vector representation of the j-th image. $\vec{t}$ is a vector of size 768, representing the text. We then compute the cosine similarity between $\vec{t}$ and $\vec{I_j}$ using Equation (6), for $j = 1 \dots N$

$$Cosine\_similarity = \frac{\vec{t} \cdot \vec{I_j}}{||\vec{t}|| \times ||\vec{I_j}||} \tag{6}$$

This gives us a vector $\vec{s}$ of size N, where s_j is a scalar corresponding to the cosine similarity between the text and the j-th image in the dataset. We sort $\vec{s}$ in descending order and analyze the indices of the sorted array. If the i-th index of the image we are analyzing is falling within the first m entries, we increase the R@m by 1. Therefore, for example, R@10 is using $m = 10$, R@5 $m = 5$, and so on. We repeat this process for all the images and compute the recall values.

The last measure we want to introduce has never been used in this context. The main idea is that if an image can be faithfully reconstructed from its description, the description is a faithful distillation of the image contents. This new evaluation method, which we call "text-to-image and comparison" (T2IC), is based on generating synthetic images from the extracted descriptions and comparing them with the original images. The higher the similarity, the more compatible and richer the image description. From each generated description, a synthetic image has been generated using a fine-tuned version of stable-diffusion-v1-4 [36]. Fine-tuning has been performed using the union of both UCM and RSICD training set images and ground-truth captions to target the generation of remote-sensing images. We used the AdamW optimizer with a learning rate of 1×10^{-5} , for 20 epochs. The measure is then articulated in two parts. The first works on couples of true and synthetic images, extracting image representations using InceptionV3 [37] and measuring the cosine similarity between them as follows:

$$T2IC = \frac{1}{N} \sum_{k=1}^{N} cosinesim(\vec{i_k}, \vec{\hat{i}_k}) \tag{7}$$

where $\vec{i_k}$ is the n-dimensional feature vector extracted from the $k-th$ original image I_k and $\vec{\hat{i}_k}$ is the n-dimensional feature vector extracted from the $k-th$ synthetic image $\hat{I}_k$. The other is the FID score, a widely used score to compare image generation systems. FID directly compares an entire dataset of original images with a dataset of synthetic images. It does so by extracting image representations with InceptionV3 and fitting two Gaussian distributions on top of the extracted feature, calculating the mean and the covariance matrix of both the original and the synthetic image distribution. Then, the FID score is calculated as the distance between the two Gaussians, measured as:

$$FID = ||\vec{m} - \hat{\vec{m}}||_2^2 + Tr(\mathbf{C} + \hat{\mathbf{C}} - 2(\mathbf{C}\hat{\mathbf{C}})^{\frac{1}{2}}) \tag{8}$$

where $\vec{m}$ and $\hat{\vec{m}}$ are the $2048 - mean$ vector for the distributions of original and synthetic images, respectively, while $\mathbf{C}$ and $\hat{\mathbf{C}}$ are the covariance matrices estimated for the distribution of original and synthetic images, respectively. The lower the FID score, the more the two distributions are similar, hence the synthetic images are similar to the original ones.

2.5. Prompts

Large language models have been shown to have great capability of prompt-following. Prompts are pieces of text that are placed before the request to direct the model to satisfy the request. As both Blip-2 and ChatGPT employ large language models, authors in [27] used specific prompts for both, placing them before question–answer generation.

2.5.1. ChatGPT Prompts

1. I have an image. Ask me questions about the content of this image. Carefully asking me informative questions to maximize your information about this image content. Each time ask one question only without giving an answer. Avoid asking yes/no questions. I'll put my answer beginning with "Answer":
2. Next Question. Avoid asking yes/no questions. Question:
3. Now summarize the information you get in a few sentences. Ignore the questions with answers no or not sure. Don't add information. Don't miss information. Summary:

Prompt (1) is used at the beginning of the dialogue to provide information to ChatGPT about what the user wants and the format in which the user wants the questions to be generated. Prompt (2) is to gather follow-up questions, and it is placed after the previous dialogue history to ask the model to provide another question in the correct format. Prompt (3) is used at the end of the dialogue to ask the model to summarize the dialogue history in a descriptive paragraph.

2.5.2. Blip-2 Prompts

1. Answer given questions. If you are not sure about the answer, say you don't know honestly. Don't imagine any contents that are not in the image.
2. Answer:

Prompt (1) is used to condition Blip-2 to output an answer only if it is sure that it is the correct answer. In [27], the authors show that if this prompt is not used, the model tends to hallucinate concepts more often. Prompt (2) is used for follow-up answers.

3. Results

The results will be split by dataset. We will report quantitative results for each metric. At the end of the section, some qualitative results are reported to further demonstrate the capability of the proposed pipeline.

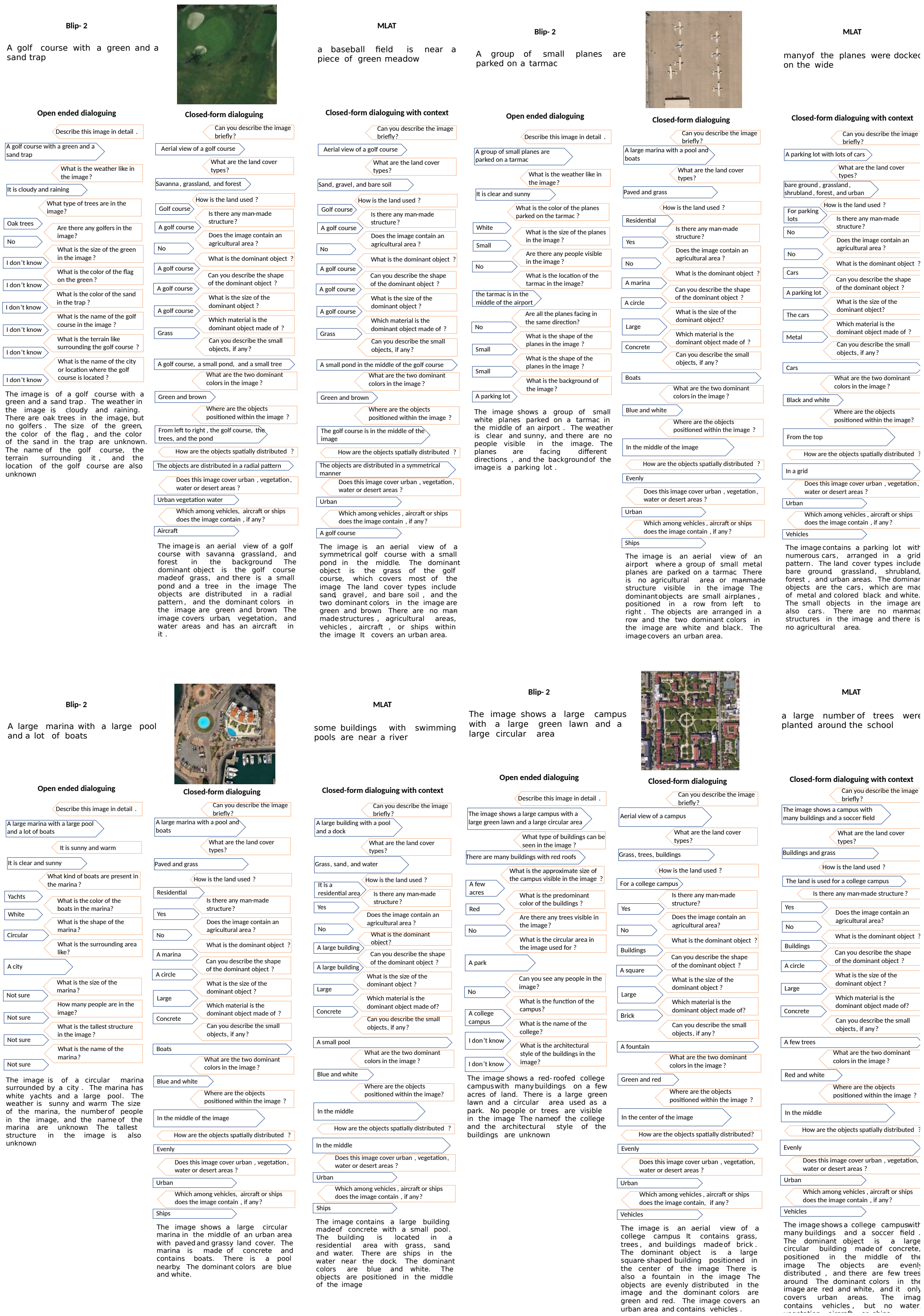
In Tables 1 and 2, the quantitative results of the three solutions described previously, namely open-ended dialogue (OED), closed-form dialogue (CFD), and closed-form dialogue with context (CFD-C), are reported. We compare the performance obtained using each generated summary by OED, CFD, and CFD-C with the results obtained using the initial answer of Blip-2. In this way, we can assess if dialoguing successfully adds value to the initial description, which is a caption-style piece of text. Inspecting the values of each of the defined metrics, we can see a trend, in which the dialogue obtained by CFD achieves the highest CLIPScore, while dialogue obtained with OED achieves the highest recall scores. This can in part be explained by the different approaches taken by OED and CFD. Indeed, in CFD, the predefined set of questions is not affected by a wrong initial description, because the question generation simply does not rely on it. For this reason, we think that summaries generated by CFD are more reliable than summaries generated with OED in situations in which the first description is inaccurate. On the other hand, the recall measure expects that the summary is more targeted because the more the summary is targeted and not general, the more an image can be differentiated from all the other images. OED obtains higher scores on the recall metrics because it does not rely on a predefined template of questions and has more possibility to target the dialogue to the input image under consideration. Lastly, CFD-C obtains the lowest recall scores, while it places in the middle between CFD and OED according to the CLIPScore. It seems that the context embedded in CFD-C is not able to improve the performance. The training using weak annotations is probably not effective since errors produced in CFD are replicated by CFD-C. We expect that with correct annotations, CFD-C can achieve higher performance by leveraging previous questions and answers to predict the next answer. Some qualitative results for both datasets are shown in Figures 7 and 8, where it can be noticed how the dialogues in the OED method are more focused, with questions that in some cases tend to be too specific and thus receive undefined answers. The CFD and CFD-C methods have more general questions and receive more general answers. In the third example, the network hallucinates some aircrafts, highlighting the necessity of a correct and coherent predefined set of questions. The T2IC score between the original and the images generated from the descriptions depict OED as the best in terms of FID score, which suggests that summaries generated through dialogue are indirectly better than the others since they allow stable-diffusion-v1-4 to reconstruct images that are more similar to the originals. Based on the T2IC score, on the RSICD dataset, it is not possible to identify the best method, with OED and CFD sharing similar scores. We believe that this can be a more truthful conclusion since RSICD is larger than UCM, and so the statistics are more reliable. Despite being very promising, this measure needs further work to be more reliable in this context. Indeed, better image generation algorithms must be used, targeted to the remote sensing scenario. We expect that leveraging image generation from rich and detailed descriptions can provide the foundation for a wider use of this measure in this context. Some examples of real and synthetic images can be found in Figure 9. It can be seen how summaries generated through dialoguing can collect more information about color, number of objects, and spatial distribution.

Table 1. Results on UCM dataset. Bold indicates the best result.

		Blip-2	**MLAT**	**Summary**		
				OED	**CFD**	**CFD-C**
Clipscore		65.2	69.0	66.1	**72.7**	67.8
ViT-L/14	R@1	21.9	4.8	**25.2**	22.9	9.0
	R@5	59.5	20.0	**61.0**	59.5	36.2
	R@10	83.8	37.1	**84.8**	81.4	63.3
T2IC	CosineSim	0.59	0.59	**0.60**	0.59	0.58
	FID	260.68	241.40	**238.71**	248.33	250.28

Table 2. Results on RSICD dataset. Bold indicates the best result.

		Blip-2	**MLAT**	**Summary**		
				OED	**CFD**	**CFD-C**
Clipscore		65.2	71.1	66.1	**72.7**	67.8
ViT-L/14	R@1	9.2	2.56	**11.0**	6.7	3.9
	R@5	29.4	12.53	**32.7**	24.4	17.2
	R@10	44.3	21.68	**49.3**	38.6	28.5
T2IC	CosineSim	0.62	0.62	**0.63**	**0.63**	0.62
	FID	140.02	151.23	**122.03**	122.89	128.65



Blip- 2

A golf course with a green and a sand trap

MLAT

a baseball field is near a piece of green meadow

Open ended dialoguing

Describe this image in detail .
A golf course with a green and a sand trap
What is the weather like in the image?
It is cloudy and raining
What type of trees are in the image?
Oak trees
Are there any golfers in the image?
No
What is the size of the green in the image?
I don't know
What is the color of the flag on the green?
I don't know
What is the color of the sand in the trap ?
I don't know
What is the name of the golf course in the image ?
I don't know
What is the terrain like surrounding the golf course ?
I don't know
What is the name of the city or location where the golf course is located ?
I don't know

The image is of a golf course with a green and a sand trap. The weather in the image is cloudy and raining. There are oak trees in the image, but no golfers. The size of the green, the color of the flag, and the color of the sand in the trap are unknown. The name of the golf course, the terrain surrounding it, and the location of the golf course are also unknown

Closed-form dialoguing

Can you describe the image briefly?
Aerial view of a golf course
What are the land cover types?
Savanna, grassland, and forest
How is the land used ?
Golf course
Is there any man-made structure?
A golf course
Does the image contain an agricultural area ?
No
What is the dominant object ?
A golf course
Can you describe the shape of the dominant object ?
A golf course
What is the size of the dominant object ?
A golf course
Which material is the dominant object made of ?
Grass
Can you describe the small objects, if any?
A golf course, a small pond, and a small tree
What are the two dominant colors in the image ?
Green and brown
Where are the objects positioned within the image ?
From left to right, the golf course, the trees, and the pond
How are the objects spatially distributed ?
The objects are distributed in a radial pattern
Does this image cover urban, vegetation, water or desert areas ?
Urban vegetation water
Which among vehicles, aircraft or ships does the image contain, if any?
Aircraft

The image is an aerial view of a golf course with savanna, grassland, and forest in the background The dominant object is the golf course made of grass, and there is a small pond and a tree in the image The objects are distributed in a radial pattern, and the dominant colors in the image are green and brown The image covers urban, vegetation, and water areas and has an aircraft in it .

Closed-form dialoguing with context

Can you describe the image briefly?
Aerial view of a golf course
What are the land cover types?
Sand, gravel, and bare soil
How is the land used ?
Golf course
Is there any man-made structure?
A golf course
Does the image contain an agricultural area ?
No
What is the dominant object ?
A golf course
Can you describe the shape of the dominant object ?
A golf course
What is the size of the dominant object ?
A golf course
Which material is the dominant object made of ?
Grass
Can you describe the small objects, if any?
A small pond in the middle of the golf course
What are the two dominant colors in the image ?
Green and brown
Where are the objects positioned within the image ?
The golf course is in the middle of the image
How are the objects spatially distributed ?
The objects are distributed in a symmetrical manner
Does this image cover urban, vegetation, water or desert areas ?
Urban
Which among vehicles, aircraft or ships does the image contain, if any?
A golf course

The image is an aerial view of a symmetrical golf course with a small pond in the middle. The dominant object is the grass of the golf course, which covers most of the image The land cover types include sand, gravel, and bare soil, and the two dominant colors in the image are green and brown There are no man made structures, agricultural areas, vehicles, aircraft, or ships within the image It covers an urban area.

Blip- 2

A group of small planes are parked on a tarmac

MLAT

manyof the planes were docked on the wide

Open ended dialoguing

Describe this image in detail .
A group of small planes are parked on a tarmac
What is the weather like in the image?
It is clear and sunny
What is the color of the planes parked on the tarmac ?
White
What is the size of the planes in the image ?
Small
Are there any people visible in the image ?
No
What is the location of the tarmac in the image?
the tarmac is in the middle of the airport
Are all the planes facing in the same direction?
No
What is the shape of the planes in the image ?
Small
What is the shape of the planes in the image ?
Small
What is the background of the image?
A parking lot

The image shows a group of small white planes parked on a tarmac in the middle of an airport . The weather is clear and sunny, and there are no people visible in the image. The planes are facing different directions , and the background of the image is a parking lot .

Closed-form dialoguing

Can you describe the image briefly?
A large marina with a pool and boats
What are the land cover types?
Paved and grass
How is the land used ?
Residential
Is there any man-made structure?
Yes
Does the image contain an agricultural area ?
No
What is the dominant object ?
A marina
Can you describe the shape of the dominant object ?
A circle
What is the size of the dominant object?
Large
Which material is the dominant object made of ?
Concrete
Can you describe the small objects, if any?
Boats
What are the two dominant colors in the image ?
Blue and white
Where are the objects positioned within the image ?
In the middle of the image
How are the objects spatially distributed ?
Evenly
Does this image cover urban, vegetation, water or desert areas ?
Urban
Which among vehicles, aircraft or ships does the image contain, if any?
Ships

The image is an aerial view of an airport where a group of small metal planes are parked on a tarmac There is no agricultural area or manmade structure visible in the image The dominant objects are small airplanes, positioned in a row from left to right . The objects are arranged in a row and the two dominant colors in the image are white and black. The image covers an urban area.

Closed-form dialoguing with context

Can you describe the image briefly?
A parking lot with lots of cars
What are the land cover types?
bare ground, grassland, shrubland, forest, and urban
How is the land used ?
For parking lots
Is there any man-made structure?
No
Does the image contain an agricultural area ?
No
What is the dominant object ?
Cars
Can you describe the shape of the dominant object ?
A parking lot
What is the size of the dominant object?
The cars
Which material is the dominant object made of ?
Metal
Can you describe the small objects, if any?
Cars
What are the two dominant colors in the image ?
Black and white
Where are the objects positioned within the image?
From the top
How are the objects spatially distributed ?
In a grid
Does this image cover urban, vegetation, water or desert areas ?
Urban
Which among vehicles, aircraft or ships does the image contain, if any?
Vehicles

The image contains a parking lot with numerous cars, arranged in a grid pattern. The land cover types include bare ground, grassland, shrubland, forest, and urban areas. The dominant objects are the cars, which are made of metal and colored black and white. The small objects in the image are also cars. There are no manmade structures in the image, and there is no agricultural area.

Blip- 2

A large marina with a large pool and a lot of boats

MLAT

some buildings with swimming pools are near a river

Open ended dialoguing

Describe this image in detail .
A large marina with a large pool and a lot of boats
It is sunny and warm
It is clear and sunny
What kind of boats are present in the marina ?
Yachts
What is the color of the boats in the marina?
White
What is the shape of the marina ?
Circular
What is the surrounding area like?
A city
What is the size of the marina ?
Not sure
How many people are in the image?
Not sure
What is the tallest structure in the image ?
Not sure
What is the name of the marina ?
Not sure

The image is of a circular marina surrounded by a city . The marina has white yachts and a large pool. The weather is sunny and warm The size of the marina, the number of people in the image, and the name of the marina are unknown The tallest structure in the image is also unknown

Closed-form dialoguing

Can you describe the image briefly?
A large marina with a pool and boats
What are the land cover types?
Paved and grass
How is the land used ?
Residential
Is there any man-made structure?
Yes
Does the image contain an agricultural area ?
No
What is the dominant object ?
A marina
Can you describe the shape of the dominant object ?
A circle
What is the size of the dominant object ?
Large
Which material is the dominant object made of ?
Concrete
Can you describe the small objects, if any?
Boats
What are the two dominant colors in the image ?
Blue and white
Where are the objects positioned within the image ?
In the middle of the image
How are the objects spatially distributed ?
Evenly
Does this image cover urban, vegetation, water or desert areas ?
Urban
Which among vehicles, aircraft or ships does the image contain, if any?
Ships

The image shows a large circular marina in the middle of an urban area with paved and grassy land cover. The marina is made of concrete and contains boats. There is a pool nearby. The dominant colors are blue and white.

Closed-form dialoguing with context

Can you describe the image briefly?
A large building with a pool and a dock
What are the land cover types?
Grass, sand, and water
How is the land used ?
It is a residential area
Is there any man-made structure?
Yes
Does the image contain an agricultural area ?
No
What is the dominant object?
A large building
Can you describe the shape of the dominant object ?
A large building
What is the size of the dominant object ?
Large
Which material is the dominant object made of?
Concrete
Can you describe the small objects, if any?
A small pool
What are the two dominant colors in the image ?
Blue and white
Where are the objects positioned within the image?
In the middle
How are the objects spatially distributed ?
In the middle
Does this image cover urban, vegetation, water or desert areas ?
Urban
Which among vehicles, aircraft or ships does the image contain, if any?
Ships

The image contains a large building made of concrete with a small pool. The building is located in a residential area with grass, sand, and water. There are ships in the water near the dock The dominant colors are blue and white. The objects are positioned in the middle of the image

Blip- 2

The image shows a large campus with a large green lawn and a large circular area

MLAT

a large number of trees were planted around the school

Open ended dialoguing

Describe this image in detail .
The image shows a large campus with a large green lawn and a large circular area
What type of buildings can be seen in the image ?
There are many buildings with red roofs
What is the approximate size of the campus visible in the image ?
A few acres
What is the predominant color of the buildings ?
Red
Are there any trees visible in the image?
No
What is the circular area in the image used for ?
A park
Can you see any people in the image?
No
What is the function of the campus?
A college campus
What is the name of the college?
I don't know
What is the architectural style of the buildings in the image?
I don't know

The image shows a red-roofed college campus with many buildings on a few acres of land. There is a large green lawn and a circular area used as a park. No people or trees are visible in the image The name of the college and the architectural style of the buildings are unknown

Closed-form dialoguing

Can you describe the image briefly?
Aerial view of a campus
What are the land cover types?
Grass, trees, buildings
How is the land used ?
For a college campus
Is there any man-made structure?
Yes
Does the image contain an agricultural area?
No
What is the dominant object ?
Buildings
Can you describe the shape of the dominant object ?
A square
What is the size of the dominant object ?
Large
Which material is the dominant object made of?
Brick
Can you describe the small objects, if any?
A fountain
What are the two dominant colors in the image ?
Green and red
Where are the objects positioned within the image ?
In the center of the image
How are the objects spatially distributed?
Evenly
Does this image cover urban, vegetation, water or desert areas ?
Urban
Which among vehicles, aircraft or ships does the image contain, if any?
Vehicles

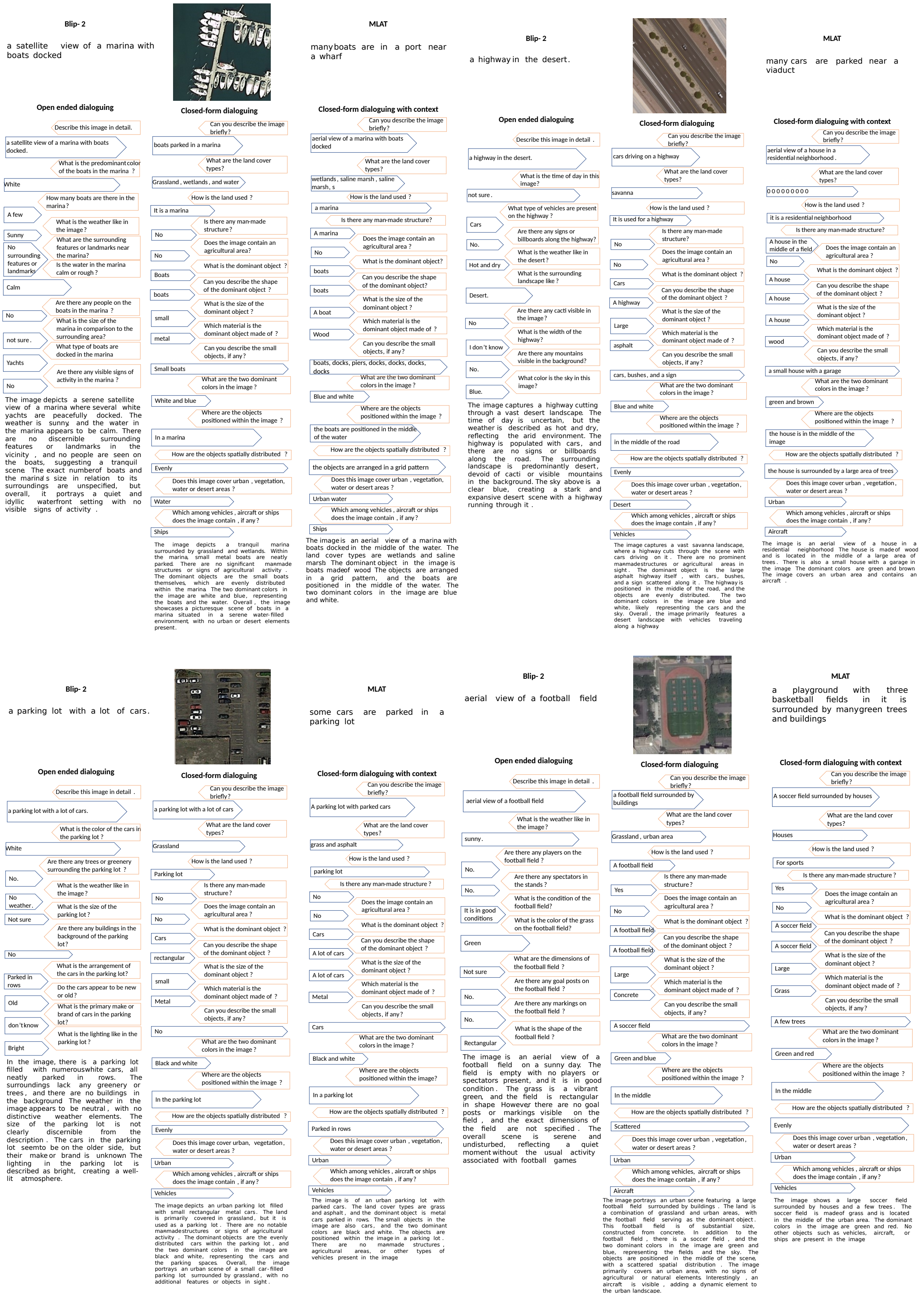
The image is an aerial view of a college campus It contains grass, trees, and buildings made of brick. The dominant object is a large square-shaped building positioned in the center of the image There is also a fountain in the image The objects are evenly distributed in the image, and the dominant colors are green and red. The image covers an urban area and contains vehicles .

Closed-form dialoguing with context

Can you describe the image briefly?
The image shows a campus with many buildings and a soccer field
What are the land cover types?
Buildings and grass
How is the land used ?
The land is used for a college campus
Is there any man-made structure ?
Yes
Does the image contain an agricultural area?
No
What is the dominant object ?
Buildings
Can you describe the shape of the dominant object ?
A circle
What is the size of the dominant object ?
Large
Which material is the dominant object made of?
Concrete
Can you describe the small objects, if any?
A few trees
What are the two dominant colors in the image ?
Red and white
Where are the objects positioned within the image ?
In the middle
How are the objects spatially distributed ?
Evenly
Does this image cover urban, vegetation, water or desert areas ?
Urban
Which among vehicles, aircraft or ships does the image contain, if any?
Vehicles

The image shows a college campus with many buildings and a soccer field . The dominant object is a large circular building made of concrete, positioned in the middle of the image The objects are evenly distributed, and there are few trees around The dominant colors in the image are red and white, and it only covers urban areas. The image contains vehicles, but no water,

Figure 7. Examples of dialogues and summaries on RSICD images (**top row**) and UCM-Captions images (**bottom row**).



Blip- 2

a satellite view of a marina with boats docked

MLAT

many boats are in a port near a wharf

Open ended dialoguing

Describe this image in detail.

a satellite view of a marina with boats docked.

What is the predominant color of the boats in the marina ?

White

How many boats are there in the marina?

A few

What is the weather like in the image?

Sunny

What are the surrounding features or landmarks near the marina?

No surrounding features or landmarks

Is the water in the marina calm or rough ?

Calm

Are there any people on the boats in the marina ?

No

What is the size of the marina in comparison to the surrounding area?

not sure.

What type of boats are docked in the marina

Yachts

Are there any visible signs of activity in the marina ?

No

The image depicts a serene satellite view of a marina where several white yachts are peacefully docked. The weather is sunny, and the water in the marina appears to be calm. There are no discernible surrounding features or landmarks in the vicinity , and no people are seen on the boats, suggesting a tranquil scene. The exact numberof boats and the marina's size in relation to its surroundings are unspecified, but overall, it portrays a quiet and idyllic waterfront setting with no visible signs of activity .

Closed-form dialoguing

Can you describe the image briefly?

boats parked in a marina

What are the land cover types?

Grassland , wetlands , and water

How is the land used ?

It is a marina

Is there any man-made structure?

No

Does the image contain an agricultural area?

No

What is the dominant object ?

Boats

Can you describe the shape of the dominant object ?

boats

What is the size of the dominant object ?

small

Which material is the dominant object made of ?

metal

Can you describe the small objects, if any?

Small boats

What are the two dominant colors in the image ?

White and blue

Where are the objects positioned within the image ?

In a marina

How are the objects spatially distributed ?

Evenly

Does this image cover urban , vegetation, water or desert areas ?

Water

Which among vehicles , aircraft or ships does the image contain , if any?

Ships

The image depicts a tranquil marina surrounded by grassland and wetlands. Within the marina, small metal boats are neatly parked. There are no significant manmade structures or signs of agricultural activity . The dominant objects are the small boats themselves, which are evenly distributed within the marina. The two dominant colors in the image are white and blue, representing the boats and the water. Overall , the image showcases a picturesque scene of boats in a marina situated in a serene water-filled environment, with no urban or desert elements present.

Closed-form dialoguing with context

Can you describe the image briefly?

aerial view of a marina with boats docked

What are the land cover types?

wetlands , saline marsh , saline marsh, s

How is the land used ?

a marina

Is there any man-made structure?

A marina

Does the image contain an agricultural area ?

No

What is the dominant object?

boats

Can you describe the shape of the dominant object?

boats

What is the size of the dominant object ?

A boat

Which material is the dominant object made of ?

Wood

Can you describe the small objects, if any?

boats, docks, piers, docks, docks, docks, docks

What are the two dominant colors in the image ?

Blue and white

Where are the objects positioned within the image ?

the boats are positioned in the middle of the water

How are the objects spatially distributed ?

the objects are arranged in a grid pattern

Does this image cover urban , vegetation, water or desert areas ?

Urban water

Which among vehicles , aircraft or ships does the image contain , if any?

Ships

The image is an aerial view of a marina with boats docked in the middle of the water. The land cover types are wetlands and saline marsh. The dominant object in the image is boats madeof wood The objects are arranged in a grid pattern, and the boats are positioned in the middle of the water. The two dominant colors in the image are blue and white.

Blip- 2

a highway in the desert.

MLAT

many cars are parked near a viaduct

Open ended dialoguing

Describe this image in detail .

a highway in the desert.

What is the time of day in this image?

not sure.

What type of vehicles are present on the highway ?

Cars

Are there any signs or billboards along the highway?

No.

What is the weather like in the desert?

Hot and dry

What is the surrounding landscape like ?

Desert.

Are there any cacti visible in the image?

No

What is the width of the highway?

I don't know

Are there any mountains visible in the background?

No.

What color is the sky in this image?

Blue.

The image captures a highway cutting through a vast desert landscape. The time of day is uncertain, but the weather is described as hot and dry, reflecting the arid environment. The highway is populated with cars, and there are no signs or billboards along the road. The surrounding landscape is predominantly desert, devoid of cacti or visible mountains in the background. The sky above is a clear blue, creating a stark and expansive desert scene with a highway running through it .

Closed-form dialoguing

Can you describe the image briefly?

cars driving on a highway

What are the land cover types?

savanna

How is the land used ?

It is used for a highway

Is there any man-made structure?

No

Does the image contain an agricultural area ?

No

What is the dominant object ?

Cars

Can you describe the shape of the dominant object ?

A highway

What is the size of the dominant object ?

Large

Which material is the dominant object made of ?

asphalt

Can you describe the small objects, if any?

cars, bushes, and a sign

What are the two dominant colors in the image ?

Blue and white

Where are the objects positioned within the image ?

in the middle of the road

How are the objects spatially distributed ?

Evenly

Does this image cover urban , vegetation, water or desert areas ?

Desert

Which among vehicles , aircraft or ships does the image contain , if any?

Vehicles

The image captures a vast savanna landscape, where a highway cuts through the scene with cars driving on it . There are no prominent manmadestructures or agricultural areas in sight . The dominant object is the large asphalt highway itself , with cars, bushes, and a sign scattered along it . The highway is positioned in the middle of the road, and the objects are evenly distributed. The two dominant colors in the image are blue and white, likely representing the cars and the sky. Overall , the image primarily features a desert landscape with vehicles traveling along a highway

Closed-form dialoguing with context

Can you describe the image briefly?

aerial view of a house in a residential neighborhood .

What are the land cover types?

000000000

How is the land used ?

it is a residential neighborhood

Is there any man-made structure?

A house in the middle of a field

Does the image contain an agricultural area ?

No

What is the dominant object ?

A house

Can you describe the shape of the dominant object ?

A house

What is the size of the dominant object ?

A house

Which material is the dominant object made of ?

wood

Can you describe the small objects, if any?

a small house with a garage

What are the two dominant colors in the image ?

green and brown

Where are the objects positioned within the image ?

the house is in the middle of the image

How are the objects spatially distributed ?

the house is surrounded by a large area of trees

Does this image cover urban , vegetation, water or desert areas ?

Urban

Which among vehicles , aircraft or ships does the image contain , if any?

Aircraft

The image is an aerial view of a house in a residential neighborhood The house is made of wood and is located in the middle of a large area of trees. There is also a small house with a garage in the image The dominant colors are green and brown The image covers an urban area and contains an aircraft .

Blip- 2

a parking lot with a lot of cars.

MLAT

some cars are parked in a parking lot

Open ended dialoguing

Describe this image in detail .

a parking lot with a lot of cars.

What is the color of the cars in the parking lot ?

White

Are there any trees or greenery surrounding the parking lot ?

No.

What is the weather like in the image?

No weather.

What is the size of the parking lot ?

Not sure

Are there any buildings in the background of the parking lot?

No

What is the arrangement of the cars in the parking lot?

Parked in rows

Do the cars appear to be new or old?

Old

What is the primary make or brand of cars in the parking lot?

don't know

What is the lighting like in the parking lot ?

Bright

In the image, there is a parking lot filled with numerouswhite cars, all neatly parked in rows. The surroundings lack any greenery or trees, and there are no buildings in the background The weather in the image appears to be neutral, with no distinctive weather elements. The size of the parking lot is not clearly discernible from the description. The cars in the parking lot seemto be on the older side, but their make or brand is unknown The lighting in the parking lot is described as bright, creating a well-lit atmosphere.

Closed-form dialoguing

Can you describe the image briefly?

a parking lot with a lot of cars

What are the land cover types?

Grassland

How is the land used ?

Parking lot

Is there any man-made structure?

No

Does the image contain an agricultural area ?

No

What is the dominant object ?

Cars

Can you describe the shape of the dominant object ?

rectangular

What is the size of the dominant object ?

small

Which material is the dominant object made of ?

Metal

Can you describe the small objects, if any?

No

What are the two dominant colors in the image ?

Black and white

Where are the objects positioned within the image ?

In the parking lot

How are the objects spatially distributed ?

Evenly

Does this image cover urban, vegetation, water or desert areas ?

Urban

Which among vehicles , aircraft or ships does the image contain , if any?

Vehicles

The image depicts an urban parking lot filled with small rectangular metal cars. The land is primarily covered in grassland, but it is used as a parking lot . There are no notable manmadestructures or signs of agricultural activity . The dominant objects are the evenly distributed cars within the parking lot , and the two dominant colors in the image are black and white, representing the cars and the parking spaces. Overall, the image portrays an urban scene of a small car-filled parking lot surrounded by grassland, with no additional features or objects in sight .

Closed-form dialoguing with context

Can you describe the image briefly?

A parking lot with parked cars

What are the land cover types?

grass and asphalt

How is the land used ?

parking lot

Is there any man-made structure ?

No

Does the image contain an agricultural area ?

No

What is the dominant object ?

Cars

Can you describe the shape of the dominant object ?

A lot of cars

What is the size of the dominant object ?

A lot of cars

Which material is the dominant object made of ?

Metal

Can you describe the small objects, if any?

Cars

What are the two dominant colors in the image ?

Black and white

Where are the objects positioned within the image?

In a parking lot

How are the objects spatially distributed ?

Parked in rows

Does this image cover urban , vegetation, water or desert areas ?

Urban

Which among vehicles , aircraft or ships does the image contain , if any?

Vehicles

The image is of an urban parking lot with parked cars. The land cover types are grass and asphalt , and the dominant object is metal cars parked in rows. The small objects in the image are also cars, and the two dominant colors are black and white. The objects are positioned within the image in a parking lot . There are no manmade structures , agricultural areas, or other types of vehicles present in the image

Blip- 2

aerial view of a football field

MLAT

a playground with three basketball fields in it is surrounded by manygreen trees and buildings

Open ended dialoguing

Describe this image in detail .

aerial view of a football field

What is the weather like in the image?

sunny.

Are there any players on the football field ?

No.

Are there any spectators in the stands ?

No.

What is the condition of the football field?

It is in good conditions

What is the color of the grass on the football field?

Green

What are the dimensions of the football field ?

Not sure

Are there any goal posts on the football field ?

No.

Are there any markings on the football field ?

No.

What is the shape of the football field ?

Rectangular

The image is an aerial view of a football field on a sunny day. The field is empty with no players or spectators present, and it is in good condition. The grass is a vibrant green, and the field is rectangular in shape However there are no goal posts or markings visible on the field , and the exact dimensions of the field are not specified. The overall scene is serene and undisturbed, reflecting a quiet moment without the usual activity associated with football games

Closed-form dialoguing

Can you describe the image briefly?

a football field surrounded by buildings

What are the land cover types?

Grassland , urban area

How is the land used ?

A football field

Is there any man-made structure?

Yes

Does the image contain an agricultural area ?

No

What is the dominant object ?

A football field

Can you describe the shape of the dominant object ?

A football field

What is the size of the dominant object ?

Large

Which material is the dominant object made of ?

Concrete

Can you describe the small objects, if any?

A soccer field

What are the two dominant colors in the image ?

Green and blue

Where are the objects positioned within the image ?

In the middle

How are the objects spatially distributed ?

Scattered

Does this image cover urban , vegetation, water or desert areas ?

Urban

Which among vehicles, aircraft or ships does the image contain , if any?

Aircraft

The image portrays an urban scene featuring a large football field surrounded by buildings . The land is a combination of grassland and urban areas, with the football field serving as the dominant object. This football field is of substantial size, constructed from concrete. In addition to the football field , there is a soccer field , and the two dominant colors in the image are green and blue, representing the fields and the sky. The objects are positioned in the middle of the scene, with a scattered spatial distribution . The image primarily covers an urban area, with no signs of agricultural or natural elements. Interestingly , an aircraft is visible , adding a dynamic element to the urban landscape.

Closed-form dialoguing with context

Can you describe the image briefly?

A soccer field surrounded by houses

What are the land cover types?

Houses

How is the land used ?

For sports

Is there any man-made structure ?

Yes

Does the image contain an agricultural area ?

No

What is the dominant object ?

A soccer field

Can you describe the shape of the dominant object ?

A soccer field

What is the size of the dominant object ?

Large

Which material is the dominant object made of ?

Grass

Can you describe the small objects, if any?

A few trees

What are the two dominant colors in the image ?

Green and red

Where are the objects positioned within the image ?

In the middle

How are the objects spatially distributed ?

Evenly

Does this image cover urban , vegetation, water or desert areas ?

Urban

Which among vehicles , aircraft or ships does the image contain , if any?

Vehicles

The image shows a large soccer field surrounded by houses and a few trees. The soccer field is madeof grass and is located in the middle of the urban area. The dominant colors in the image are green and red. No other objects such as vehicles, aircraft, or ships are present in the image

Figure 8. Examples of dialogues and summaries on RSICD images (**top row**) and UCM-Captions images (**bottom row**).

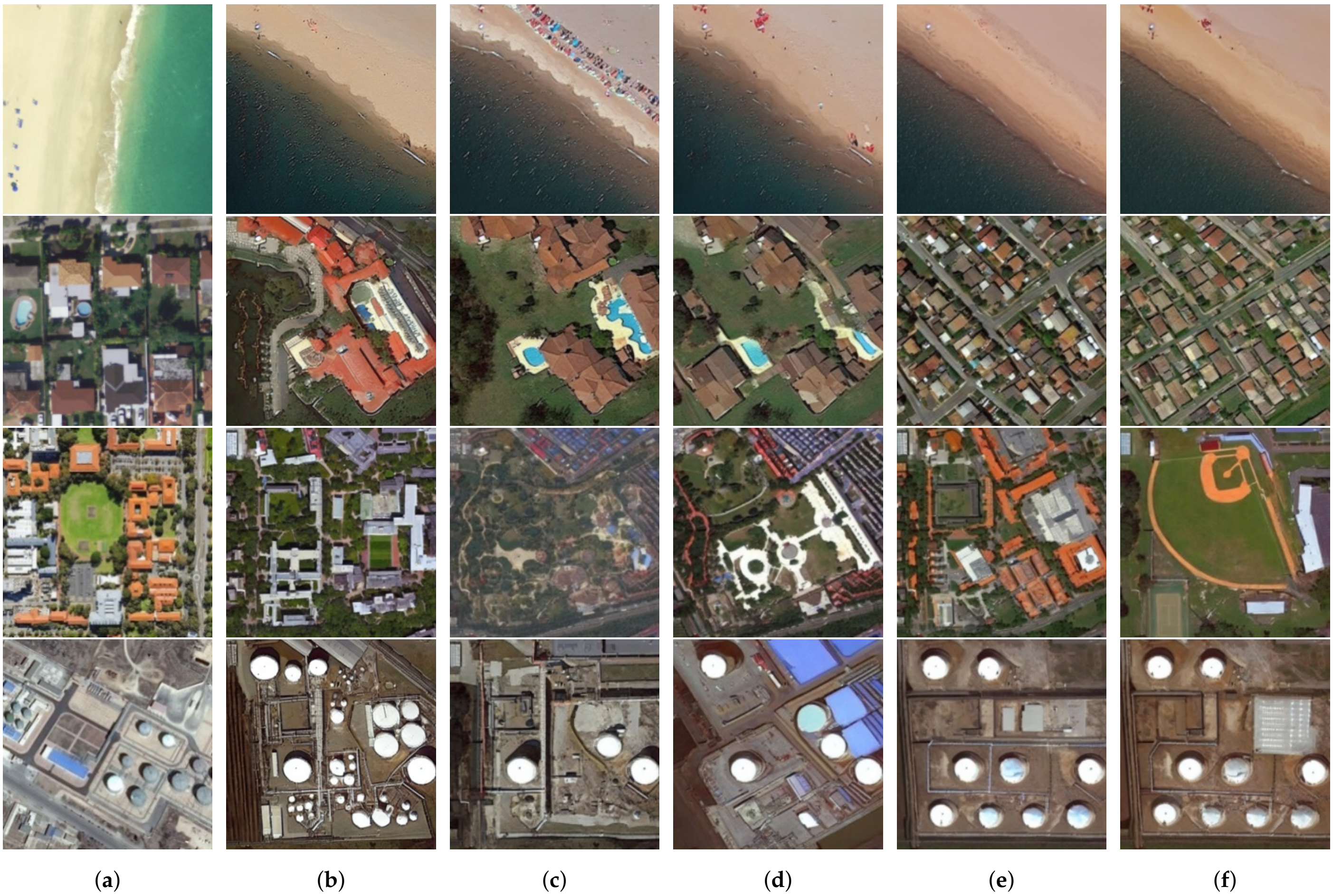


Figure 9. Original (**a**) and synthetic images generated starting from descriptions generated with different methods. (**b**) MLAT, (**c**) Blip-2, (**d**) OED, (**e**) CFD, and (**f**) CFD-C.

4. Discussion

This paper presents a new approach for remote sensing image description generation that, thanks to a dialogue between two machines, allows it to further dig information from images and convert it into a natural language. The approach is innovative with respect to traditional captioning methods, which instead produce a one-shot description that barely can capture the rich semantics of remote sensing images. Despite the promising results, several weaknesses must be addressed to build a more robust and customized pipeline.

4.1. Input Image Integration

In this work, the conditioning of the dialogue on the input image is limited to the answering process. The questioner, a large language model (ChatGPT), exclusively relies on the textual exchange to build subsequent questions. We noticed that this can severely impact the performance when the initial description from Blip-2 is inaccurate. Despite the closed-ended strategy seeming to cope better with this problem, we argue that a better integration of the visual information in the dialoguing process is of outmost importance and can improve the overall process by making the questions more targeted to the image itself.

4.2. Adaptation of Answerer and Questioner to the RS Context

Both the networks employed in OED and CFD have been used in a zero-shot fashion, leading to criticalities when the network is presented with remote sensing images. In particular, the answerer is the most critical part since it deals with the image. The questioner is not directly affected, since it relies only on the textual information to generate questions, but errors in the answerer indirectly impact the generation of subsequent questions. To alleviate this problem, remote sensing visual question answering datasets can be exploited to refine the algorithm that generates the answers. The questioner, on the other hand, can be fine tuned using ground-truth conversation, opening the path to the creation of a dataset for RS visual dialoguing. However, the expensiveness of this process creates the need to find solutions that lie somewhere between human and automated annotation, such as trying to filter the conversations created by isolating the most coherent and plausible conversations for the remote sensing domain. Lastly, we point out that our proposed CFD-C is not to be considered a fine-tuned version, since it uses the results of CFD as soft labels for training.

4.3. Removal of Uncertain Question–Answer Pairs

With the use of the uncertainty prompt in Blip-2, several questions, especially in the open-ended dialogue, receive "I don't know" or "not sure" answers. Authors in [28] adopted a specific ChatGPT prompt to try to eliminate this redundant and useless information when generating the summary. However, in our context, this approach is not effective, and we argue that authors in [28] did not notice this criticality because they received less answers of this type, being in a natural image scenario, where Blip-2 is more effective. A simpler way to deal with such a situation can be to remove all the question–answer pairs that receive uncertain answers prior to the summarization step.

4.4. Better Predefined Questions for CFD Method

We noticed that the definition of some questions for the CFD method turned out to be a weakness. This is the case, for example, of the question "Which among vehicles, aircraft or ships does the image contain, if any?", that, in situations where none of these objects are present, misleads Blip-2, leading to erroneous outputs. Specifically, for CFD, creating more specific templates can lead to more detailed descriptions, especially in situations where the user already knows which information to extract.

4.5. Evaluation Metrics and Generation of Targeted Datasets

The evaluation strategies adopted in this paper use different kinds of reference-free metrics. In general, these metrics provide indirect measures of the goodness of the generated dialogues. To fully evaluate the consistency of a dialogue directly of the generated description, it would be necessary to have ground-truth summaries generated by human operators. This could give rise to new datasets targeted to the task of richer image description through a paragraph. However, the generation of ground-truth summaries is very costly and time-consuming. As pointed out in subsection B, joint use of human operator and machine-generated data can be useful to speed up the creation of such datasets. At the same time, research on other suitable reference-free metrics must be conducted, as it offers a valuable alternative.

4.6. Customizing the Dialogue and Multimodality

Another aspect that can be worth exploring is the targeting of the dialogue on the need of the user. As explored in this work, large language models adopt prompts to focus their attention on specific outputs needed by the user. This can give rise to multimodal-driven dialoguing, in which the user is also able, through the definition of specific prompts, to direct the attention of the dialoguing process to specific aspects of the image that are worth analyzing. Finding the best way to introduce such conditioning on the dialogue can be a major advancement, since it enables users to interact with the system in a natural and engaging way through natural language. On the other hand, it can provide the user with a more useful and detailed description of the highlighted concepts. Practically, this approach can be seen as an automatic generalization of CFD, in which instead of fixing a priori the questions, the user directs the questioner in producing set of questions targeted to the need of the moment, alleviating the burden of the question definition. Moreover, one may extend the future perspective to scenarios where the user not only may interact with the dialoguing system but feeds it with other sources of information such as ancillary information from geographic information systems.

5. Conclusions

In this paper, we propose three approaches to extract richer textual descriptions of remote sensing images. Our three approaches use machine-to-machine dialogue to extract information sequentially and in small bits. Once the dialogue is concluded, its content is summarized in a paragraph, creating the final description. We evaluate the coherence of our captions using different indirect metrics, where our open-handed dialogue generally obtains the higher scores. Each of the approaches is targeted to a different use case. Where open-handed dialogue can retrieve general descriptions, in certain applications where it is possible to define a set of predefined set of questions, closed-form dialoguing can be the best choice to customize the descriptions.

Author Contributions: Conceptualization, R.R., Y.B. and F.M.; Methodology, R.R. and Y.B.; Validation, R.R. and Y.B.; Writing—original draft, R.R.; Writing—review & editing, Y.B. and F.M.; Supervision, F.M.; Project administration, F.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets used in this work are publicly available and can be found https://github.com/201528014227051/RSICD_optimal (accessed on 18 January 2024). Data are contained within the article. We release all our code https://github.com/RicRicci22/M2MVD.git (accessed on 18 January 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

R@1-5-10	Recall value considering the first match, the first five, and the first ten, respectively.
T2IC	Text-to-Image and Comparison
FID	Frechet Inception Distance
OED	Open-Ended Dialogue
CFD	Closed-Form Dialogue
CFD-C	Closed-Form Dialogue with Context

References

1. Wu, L.; Tan, X.; He, D.; Tian, F.; Qin, T.; Lai, J.; Liu, T.Y. Beyond Error Propagation in Neural Machine Translation: Characteristics of Language. *arXiv* **2018**, arXiv:1809.00120.
2. Hoxha, G.; Chouaf, S.; Melgani, F.; Smara, Y. Change Captioning: A New Paradigm for Multitemporal Remote Sensing Image Analysis. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
3. Shi, Z.; Zou, Z. Can a Machine Generate Humanlike Language Descriptions for a Remote Sensing Image? *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3623–3634. [CrossRef]
4. Qu, B.; Li, X.; Tao, D.; Lu, X. Deep semantic understanding of high resolution remote sensing image. In Proceedings of the 2016 International Conference on Computer, Information and Telecommunication Systems (CITS), Kunming, China, 6–8 July 2016 ; pp. 1–5. [CrossRef]
5. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
6. Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv* **2014**, arXiv:1406.1078.
7. Hoxha, G.; Melgani, F. A Novel SVM-Based Decoder for Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
8. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2183–2195. [CrossRef]
9. Huang, W.; Wang, Q.; Li, X. Denoising-Based Multiscale Feature Fusion for Remote Sensing Image Captioning. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 436–440. [CrossRef]
10. Zhang, X.; Wang, X.; Tang, X.; Zhou, H.; Li, C. Description Generation for Remote Sensing Images Using Attribute Attention Mechanism. *Remote Sens.* **2019**, *11*, 612. [CrossRef]

11. Zhang, Z.; Diao, W.; Zhang, W.; Yan, M.; Gao, X.; Sun, X. LAM: Remote Sensing Image Captioning with Label-Attention Mechanism. *Remote Sens.* **2019**, *11*, 2349. [CrossRef]
12. Wang, J.; Chen, Z.; Ma, A.; Zhong, Y. Capformer: Pure Transformer for Remote Sensing Image Caption. In Proceedings of the IGARSS 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022 ; pp. 7996–7999. [CrossRef]
13. Cheng, Q.; Huang, H.; Xu, Y.; Zhou, Y.; Li, H.; Wang, Z. NWPU-Captions Dataset and MLCA-Net for Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–19. [CrossRef]
14. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. VQA: Visual Question Answering. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
15. Lobry, S.; Marcos, D.; Murray, J.; Tuia, D. RSVQA: Visual Question Answering for Remote Sensing Data. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8555–8566. [CrossRef]
16. Zheng, X.; Wang, B.; Du, X.; Lu, X. Mutual Attention Inception Network for Remote Sensing Visual Question Answering. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
17. Yuan, Z.; Mou, L.; Wang, Q.; Zhu, X.X. From Easy to Hard: Learning Language-Guided Curriculum for Visual Question Answering on Remote Sensing Data. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
18. Chappuis, C.; Zermatten, V.; Lobry, S.; Le Saux, B.; Tuia, D. Prompt–RSVQA: Prompting visual context to a language model for Remote Sensing Visual Question Answering. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 19–20 June 2022; pp. 1371–1380. [CrossRef]
19. Bazi, Y.; Rahhal, M.M.A.; Mekhalfi, M.L.; Zuair, M.A.A.; Melgani, F. Bi-Modal Transformer-Based Approach for Visual Question Answering in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
20. Patil, C. Visual Question Generation: The State of the Art. *ACM Comput. Surv.* **2020**, *53*, 1–22. [CrossRef]
21. Ren, M.; Kiros, R.; Zemel, R. Exploring Models and Data for Image Question Answering. *arXiv* **2015**, arXiv:1505.020742.
22. Geman, D.; Geman, S.; Hallonquist, N.; Younes, L. Visual Turing test for computer vision systems. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 3618–3623. [CrossRef]
23. Yang, J.; Lu, J.; Lee, S.; Batra, D.; Parikh, D. Visual Curiosity: Learning to Ask Questions to Learn Visual Recognition. *arXiv* **2018**, arXiv:1810.00912.
24. Vedd, N.; Wang, Z.; Rei, M.; Miao, Y.; Specia, L. Guiding Visual Question Generation. *arXiv* **2012**, arXiv:2110.08226.
25. Jain, U.; Zhang, Z.; Schwing, A. Creativity: Generating Diverse Questions using Variational Autoencoders. *arXiv* **2017**, arXiv:1704.03493.
26. Bashmal, L.; Bazi, Y.; Melgani, F.; Ricci, R.; Al Rahhal, M.M.; Zuair, M. Visual Question Generation From Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 3279–3293. [CrossRef]
27. Zhu, D.; Chen, J.; Haydarov, K.; Shen, X.; Zhang, W.; Elhoseiny, M. ChatGPT Asks, BLIP-2 Answers: Automatic Questioning Towards Enriched Visual Descriptions. *arXiv* **2023**, arXiv:2303.06594.
28. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv* **2023**, arXiv:2301.12597.
29. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165.
30. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *arXiv* **2022**, arXiv:2203.02155.
31. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. *arXiv* **2021**, arXiv:2103.00020.
32. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* **2019**, *1*, 9.
33. Huggingface: Distilgpt2. Available online: https://huggingface.co/distilgpt2 (accessed on 14 April 2022).
34. Yang, Y.; Li, Y.; Fermuller, C.; Aloimonos, Y. Neural Self Talk: Image Understanding via Continuous Questioning and Answering. *arXiv* **2015**, arXiv:1512.03460.
35. Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R.L.; Choi, Y. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. *arXiv* **2022**, arXiv:2104.08718
36. Huggingface: CompVis/Stable-Diffusion-v1-4. Available online: https://huggingface.co/CompVis/stable-diffusion-v1-4 (accessed on 14 April 2022).
37. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. *arXiv* **2015**, arXiv:1512.00567.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.