computers

Article

SADAMB: Advancing Spatially-Aware Vision-Language Modeling Through Datasets, Metrics, and Benchmarks

Giorgos Papadopoulos, Petros Drakoulis *, Athanasios Ntovas, Alexandros Doumanoglou and Dimitris Zarpalas *

Centre for Research and Technology HELLAS (CERTH), Information Technologies Institute (ITI), 57001 Thessaloniki, Greece; giorgospap@iti.gr (G.P.); atdovas@iti.gr (A.N.); aldoum@iti.gr (A.D.)
* Correspondence: petros.drakoulis@iti.gr (P.D.); zarpalas@iti.gr (D.Z.)

Abstract:

Understanding spatial relationships between objects in images is crucial for robotic navigation, augmented reality systems, and autonomous driving applications, among others. However, existing vision-language benchmarks often overlook explicit spatial reasoning, limiting progress in this area. We attribute this limitation in part to existing open datasets and evaluation metrics, which tend to overlook spatial details. To address this gap, we make three contributions: First, we greatly extend the COCO dataset with annotations of spatial relations, providing a resource for spatially aware image captioning and visual question answering. Second, we propose a new evaluation framework encompassing metrics that assess image captions' spatial accuracy at both the sentence and dataset levels. And third, we conduct a benchmark study of various vision encoder–text decoder transformer architectures for image captioning using the introduced dataset and metrics. Results reveal that current models capture spatial information only partially, underscoring the challenges of spatially grounded caption generation.

Keywords: vision-language modeling; spatial relations; spatial grounding; spatial image captioning; spatial visual question answering; dataset; metrics; benchmark

Academic Editor: Zheng Wang

Received: 28 August 2025
Revised: 16 September 2025
Accepted: 22 September 2025
Published: 29 September 2025

Citation: Papadopoulos, G.; Drakoulis, P.; Ntovas, A.; Doumanoglou, A.; Zarpalas, D. SADAMB: Advancing Spatially-Aware Vision-Language Modeling Through Datasets, Metrics, and Benchmarks. *Computers* **2025**, *14*, 413. https://doi.org/10.3390/computers14100413

1. Introduction

Image captioning (IC) [1] is a fundamental task in computer vision and natural language processing (NLP) [2–4] that aims to generate natural language descriptions of images. While traditional models have succeeded in capturing general image content, they often overlook fine-grained spatial relationships between objects—descriptions such as "to the left of" or "behind of", which are essential in domains like autonomous driving [5], augmented reality systems [6], robotics [7], and medical imaging [8]. Recent advances in vision-language models, including Vision Transformers [9] combined with text decoders like GPT [10], have enhanced the ability to model complex contextual relationships. However, deeper understandings of spatial semantics remain largely unexplored.

A key limitation is that most existing datasets and benchmarks do not provide explicit spatial annotations or descriptions, preventing models from learning or generalizing spatial reasoning effectively. Consequently, even state-of-the-art captioning systems tend to produce generic sentences focused on object or action recognition while failing to articulate spatial context. Furthermore, widely used evaluation metrics such as BLEU [11], METEOR [12], and ROUGE [13] only indirectly account for spatial accuracy, through their impact on the lexical and grammatical coherence of the generated sentences. This creates a significant gap in our ability to evaluate and improve models for spatially grounded image captioning.

To address these limitations, we make three key contributions. First, we greatly extend the COCO [14] dataset by incorporating spatial descriptions into the image captions. In addition to the enriched captions, we also generate question–answer pairs derived directly from these spatial descriptions, to also enable the spatially aware visual question answering (VQA) [15] task. Furthermore, we provide depth maps, object-level metadata, and binary masks for all detected objects, providing richer contextual and geometric information. Second, we introduce a new spatial metric designed to evaluate the presence and accuracy of spatial relationships in image captions. This metric complements existing evaluation tools by providing insight into a model's spatial reasoning [16] capabilities. Together, these contributions allow for a comprehensive benchmark study (the third contribution) for evaluating vision-encoder text-decoder architectures [17], pushing the boundaries of image captioning toward spatial grounding and understanding. Through this work, our goal is to encourage the development of models that go beyond object recognition and achieve deeper spatial comprehension, ultimately enabling more effective deployment in complex, real-world tasks.

2. Related Work

The concept of combining vision and language models for image captioning has been extensively explored in recent years. However, much of the focus has remained on object recognition without emphasizing the spatial relationships between objects in the scene. In this section, we will compare existing datasets, metrics, and benchmarks in the context of spatially grounded image captioning and spatial reasoning, which closely relate to our work in creating a large spatially aware dataset and introducing a new family of spatial metrics.

2.1. Existing Datasets

COCO (Common Objects in Context):

The work of Lin et al. [14] has long been a benchmark in object detection, segmentation, and captioning tasks. While it contains rich annotations for object categories and image captions, it does not explicitly capture spatial relationships between objects in a scene. Existing models trained on COCO typically focus on object identification without spatial reasoning. Our work greatly extends the COCO dataset by incorporating explicit spatial descriptions, enriching each image with annotations that describe the relative positions of objects. This makes it more suitable for spatial reasoning tasks.

Visual Genome:

Krishna et al. [18] propose a large-scale dataset with annotations at the image, object, and region levels. It also includes relationships between objects, making it fundamentally spatially aware. For example, it contains object annotations with "in front of" or "next to" descriptions. However, Visual Genome focuses on specific object–object relationships, and it does not offer comprehensive natural language descriptions of spatial contexts in an image. Our work draws inspiration from these insights but provides a much richer array of spatial descriptions.

Clevr:

Johnson et al. [19] is a dataset designed to test spatial reasoning in visual question answering. It provides synthetic images of simple 3D-scenes with a variety of spatial relations, such as "to the left of," "above," and "below." While CLEVR focuses on structured, synthetic environments, it lacks the real-world complexity of datasets like COCO. Moreover, its captions are limited to very basic spatial relations and do not reflect the more nuanced descriptions found in natural scenes. Our work goes beyond CLEVR by introducing a larger-scale, real-world dataset suitable for both visual question answering and captioning training and evaluation.

GQA:

Hudson and Manning [20] is a dataset specifically designed to test compositional generalization in visual reasoning. It provides questions about the spatial relations between objects (e.g., “Is the cup next to the plate?”), which resembles our focus on spatial relations. However, it does not provide natural language captions or detailed spatial descriptions embedded in fluent sentence form. In contrast, our dataset provides richly annotated image captions that naturally incorporate spatial relationships, while also generating VQA-compatible question–answer pairs, bridging the gap between captioning and reasoning tasks.

Spatial Commonsense:

Storks et al. [21] includes annotations about spatial relationships between objects in images, something similar to our objective. It focuses on common spatial terms (e.g., “on the left,” “behind”) but lacks depth and all the supplementary modalities present in our dataset. Our method surpasses this by offering high-resolution depth maps, segmentation masks, and object-level metadata, enabling models to reason about spatiality using both textural and geometric cues. This holistic, multi-modal approach creates a significantly more robust benchmark for developing and evaluating spatially aware vision-language models.

Visual Spatial Reasoning:

Liu et al. Liu et al. [22] present the Visual Spatial Reasoning (VSR) dataset, a large-scale resource containing over 10,000 image–text pairs that capture 66 distinct types of spatial relations, such as “under”, “in front of”, and “facing towards”. What is notable about this work is the number of spatial relations evaluated, the variability of the reference frame for inducing the spatial relations, and its comparisons against the human element. Human performance on VSR exceeds 95% while state-of-the-art models achieve only around 70%, revealing a substantial gap and underscoring the challenge of spatial reasoning for multimodal AI systems. Compared to this prominent work, while SADAMB lacks a plethora of its spatial relations and uses only a fixed frame of reference, our work uniquely includes different phraseology reserved solely for relations between similar-class objects (i.e., “in a row”, “one on top of the other”, “side by side”). As regards the number of image–text pairs, our work greatly exceeds VSR by orders of magnitude. Additionally, while VSR is formulated essentially as a binary VQA setup (where each statement about an image is evaluated for being either true of false), SADAMB provides material for fully comprehensive VQA training.

2.2. Conventional Metrics for Generic IC

Traditional metrics such as BLEU, ROUGE, and METEOR assess surface-level similarity through n-gram overlap or semantic alignment but do not explicitly assess spatial accuracy. To address this, we propose **Spatial Captioning Accuracy (SCA)**, which builds beyond existing metrics with spatial correctness checks.

BLEU:

Papineni et al. [11] has been the most widely used metric in IC tasks, primarily measuring n-gram overlap between predicted and ground truth captions. However, BLEU does not explicitly capture the truthiness of spatial relationships between objects, making it insufficient for evaluating spatially aware captioning.

ROUGE:

Lin [13] is a set of metrics that measure the overlap of n-grams, word sequences, and word pairs between predicted and reference texts. While useful for evaluating content coverage and recall, ROUGE, still, does not consider the spatial structure or relationships between objects, rendering it limited for assessing spatially grounded captions.

METEOR:

Banerjee and Lavie [12] is another common metric used for IC tasks. It improves upon both BLEU and ROUGE in the evaluation of the alignment of words by taking into account synonyms and stemming. While METEOR can incidentally evaluate semantic meaning, it also falls short when it comes to evaluating the spatial aspect of object relationships in its essence.

Despite the significant advancements in image captioning and vision-language modeling, several critical gaps remain in the research community, particularly concerning the integration of spatial reasoning in captioning and question-answering tasks. This work addresses several of these gaps by greatly enhancing an established dataset, introducing a new family of spatial metrics, and establishing a comprehensive benchmark for spatial IC (not for VQA in this iteration of the work). Below, we highlight the key gaps in the literature that this work aims to address:

2.3. Lack of Rich Spatial Descriptions in Existing Datasets

While several vision-language datasets, such as COCO, Visual Genome, and CLEVR exist that include object annotations and relationships, none of these datasets fully integrate complex, natural spatial descriptions in a way that mirrors human language. Existing datasets focus on basic object relationships, such as “next to” or “in front of,” but they often fail to capture the richness and diversity of spatial relationships that naturally occur in real-world scenarios.

Our descriptions capture not only standard object–object spatial relationships such as “above,” “to the left of,” and “behind,” but also less common configurations between similar objects, including “side by side,” “one on top of the other”, and “one in front of the other.” This addition fills a significant gap by providing a dataset that reflects the pluralism and variety of spatial reasoning needed for image-captioning tasks.

2.4. Inadequate Evaluation Metrics for Spatial Reasoning

While some metrics, such as CIDEr, consider sentence fluency and consistency, they do not directly evaluate the correctness of spatial relationships between objects, which is essential for tasks requiring spatial awareness.

Our **Spatial Captioning Accuracy (SCA)** metric provides a more comprehensive way of evaluating models that generate captions with spatial relationships. These metrics allow for the assessment of both semantic and spatial accuracy, offering a more holistic view of a model’s performance. We should add here that SCA is inspired and designed to address the peculiarities of our specific setup, which includes powerful, transformer-based, neural architectures and limited language vocabulary. Thus, while we consider it optimal for comparisons in our setup, we do not claim that it is directly applicable to every benchmark setup retaining its characteristics, especially in scenarios involving less powerful architectures or significantly more complex vocabulary.

2.5. Limited Scope of Existing Datasets for Spatially Aware VQA

While many existing image captioning datasets, such as COCO, focus on generating descriptive captions for images, they often overlook the potential for these captions to be leveraged for visual question answering tasks. VQA typically involves generating answers to natural language questions about images, yet most image captioning datasets are not explicitly structured to support VQA-style question–answer pairs.

Our dataset extends captioning into the VQA domain by generating structured question–answer pairs stemming directly from spatial captions. For example:

Generated Caption:

Caption: "a car is to the left of a person."

Generated Q & A:

Question: "where is the car?"
Answer: "to the left of the person."

This approach serves two purposes:

1. Enhances the Utilization of Spatial Captions: The spatial captions generated by this work not only describe the objects in an image, but also provide context for more complex VQA tasks that require an understanding of spatial relationships.
2. Bridges the Gap between Captioning and VQA: By generating question–answer pairs from the same spatially enriched captions, we enable a seamless connection between the two tasks, allowing for multi-task learning where models can simultaneously generate captions and answer spatially aware questions about images. This cross-task capability is critical for models that aim to tackle a broader range of vision-language tasks, making them more versatile and efficient in real-world applications.

All in all, this work provides a unique resource for training and evaluating models on both IC and VQA tasks, with a particular focus on spatial reasoning. This dual-task approach facilitates the development of vision-language models that can generate descriptive captions and answer detailed spatial questions about images, thus contributing to more robust and versatile multi-modal AI systems.

3. Dataset

3.1. Dataset Creation Process

We introduce **SADAMB**, a large, structured extension of the COCO dataset enriched with spatial relationship annotations between objects and various auxiliary information and modalities. The spatial caption generation is achieved automatically, based on object recognition and depth estimation information provided by external pre-trained models. A histogram with the count of all the final dataset's objects can be seen at Figure 1. We consider all descriptions to be relative to the viewer's point of view, and the sentence composition is guided via heuristic rules encompassing threshold values and Boolean logic applied on the various objects. For the development of these rules and the selection of the various threshold values, no structured human evaluation study was conducted. We resorted to manual examination by the authoring team of 100 randomly selected dataset images, where we tried to identify values that apparently increased the number of detected relations in the images without considerably hurting the overall accuracy of the resulted descriptions. Consequently, we do not claim that the total of the generated captions and question–answer pairs is correct; we only support that in its vastness, given some margin for error which we are unable to reliably quantify at this moment, our dataset is still very useful for spatial visual training and evaluation. Furthermore, despite our best efforts, we acknowledge that our pipeline relies heavily on the accuracy of the external models used and the overall quality of our heuristics, leading sometimes to the generation of erroneous spatial descriptions.

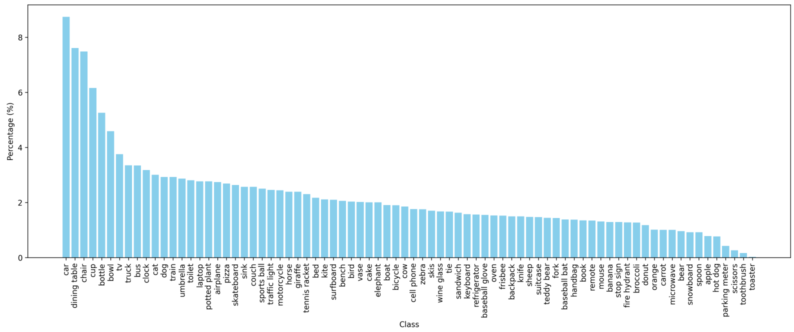


Figure 1. Distribution of object classes in the dataset, excluding the most frequent class: person.

The construction process begins with object detection using the `yolov8` [23] model. Detected objects with a confidence score below 0.45 are discarded. To ensure relevance, only objects occupying less than or equal to 90% of the image area are retained. Among these, the top four objects with the largest apparent area are selected for spatial relationship extraction. Images that contain no detected objects are discarded, while those with only a single object are denoted as "solo" and described using the template, "there is a/an <class>.".

Sentence construction follows strict grammatical and stylistic conventions: all descriptions are written entirely in lowercase, including the sentence-first letter, and each sentence terminates with a punctuation mark. Numerical values are written in letter form, and the correct article (*a* or *an*) is determined based on the phonetic sound of the object name. Some class names are also normalized for grammatical correctness, e.g., "skis" is converted to "ski" and "scissors" becomes "pair of scissors." Additionally, duplicate sentences and redundant relationship descriptions within an image are discarded.

Depth estimation is performed using the pre-trained `ZoeDepth` [24] model. Depth information is utilized alongside geometric features of the bounding boxes in a heuristic manner to extract meaningful spatial relationships. Three primary spatial relationships are considered: horizontal (*left of, right of, side by side*), vertical (*above, below, one on top of the other*), and depth-based (*in front of, behind, one in front of the other*). For horizontal relationships, object i is considered to be to the right of object j if its center is horizontally further right, the normalized center separation exceeds 0.325, vertical misalignment is minimal (less than 3.2), and their depths are close enough (within 3.8 units). Bounding boxes must indicate that object i lies entirely to the right of object j. If both objects are of the same class, the phrase includes the word "another" to reflect this. All left–right relationships for the same objects are replaced with *side by side*.

For vertical relationships, object i is said to be below object j if its vertical center lies beneath that of j, the normalized vertical separation exceeds 0.325, horizontal misalignment is below 3.2, and the depth difference is minimal. Additionally, bounding boxes must indicate that object i lies beneath j. The inverse rule applies for detecting *above* relationships, and both directions are replaced with *one on top of the other*.

Depth-based relationships are defined by comparing depth values directly. Specifically, object i is labeled as being behind object j if its depth value is greater by more than 0.8, and both horizontal and vertical misalignment is minimal (each under 3.2 in normalized units). The inverse relationship defines *in front of*, and both forms are replaced in the output as *one in front of the other*. The prevalence of each of the recognized spatial relationships in the dataset can be seen in Figure 2.

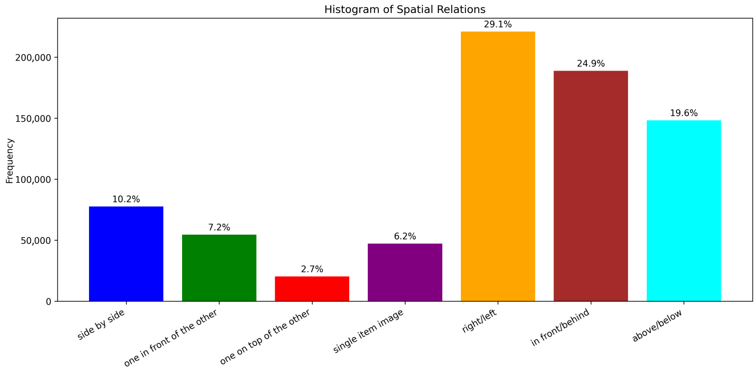


Figure 2. Histogram depicting the frequency distribution of spatial relationships in the dataset.

This carefully designed pipeline facilitates, to a good degree, the generation of accurate, disambiguated, and linguistically consistent spatial descriptions. All thresholds were chosen in an effort to minimize the presence of "phantom" relations in the dataset. Primarily, erroneous relations can arise for four reasons:

1. The `yolov8` [23] object detector provides the wrong class for an object or completely misses it.
2. The `ZoeDepth` [24] depth estimator provides depth estimations for the objects that are not compatible with their true order along the Z(depth)-axis.
3. Our heuristics that manipulate the estimated 3D bounding-boxes of the objects (calculated combining information from the two previously mentioned pre-trained models) consider relations that a human evaluator would not due to the imperfect, oversimplified nature of the logical examinations occurring.
4. There are more than four well-sized objects in the scene, while we only consider up to four. This arbitrary hard limit was enforced to prevent the creation of very long image descriptions (since we consider all vs. all objects) that would require very large token sizes to be processed, rendering a good portion of the dataset irrelevant for training models compatible with our targeted edge devices. This strategy, of course, can sometimes lead to an odd situation where not all individual objects of a class are treated (e.g., conducting bad object counting). In any case, the object count and its following distribution for all data splits can be seen in Figure 3.

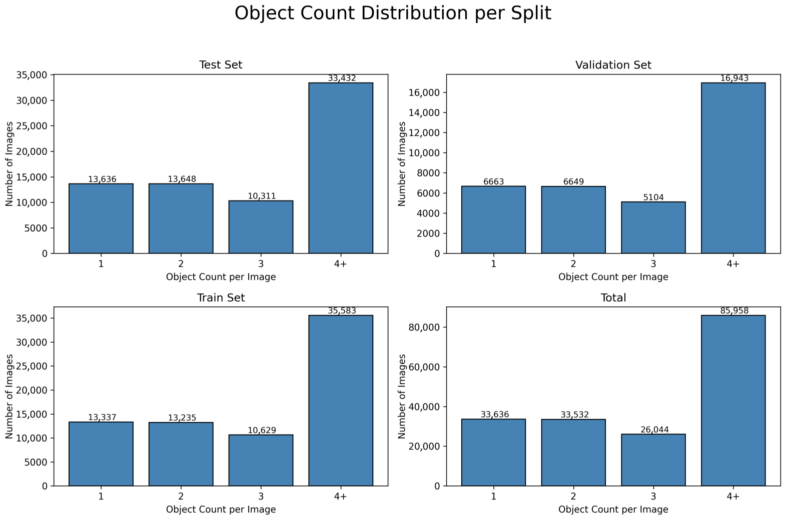


Figure 3. Distribution of object counts per image across dataset splits.

3.2. Dataset Extension for Spatial VQA

To support the task of VQA with spatial reasoning, we extended our dataset by additionally generating structured question–answer pairs, deriving from the same information that the captions do. The same object instances, as recognized by the detector in the previous IC task, and all their auxiliary information, were utilized appropriately. No humans were involved in the question–answer dataset extraction, and for each image, Q&A pairs were programmatically crafted using templates depending on the number and spatial configurations of each detected object.

All Q&A pairs were generated using consistent linguistic rules, ensuring alignment with the captioning methodology described above. This process yielded a unified dataset suitable for training and evaluating spatial models in both caption generation and visual question answering tasks (although in this work we only provide new metrics and benchmarks for the spatial captioning task).

In its totality, the produced dataset spans across 179,170 images, each one accompanied by auxiliary visual modalities (Figure 4): 758,068 extracted spatial captions and 2,394,565 spatial question–answer pairs, all split into three sub-sets (with respect to the original COCO dataset) for training, validation, and testing purposes. It should be noted that 25,551 images from the original COCO dataset were discarded because they did not contain any recognizable objects, as per our pipeline's heuristics. A summary of the characteristics of our dataset can be seen in Table 1, accompanied with some indicative samples in Figure 5. For transparency and completeness, some erroneous samples can also be seen in Figure 6.

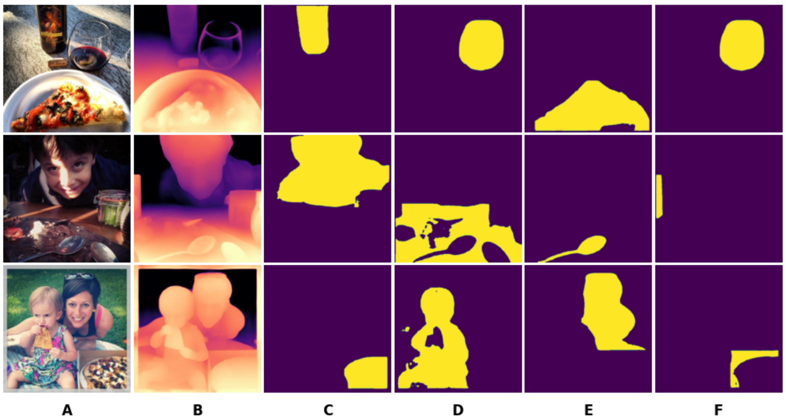
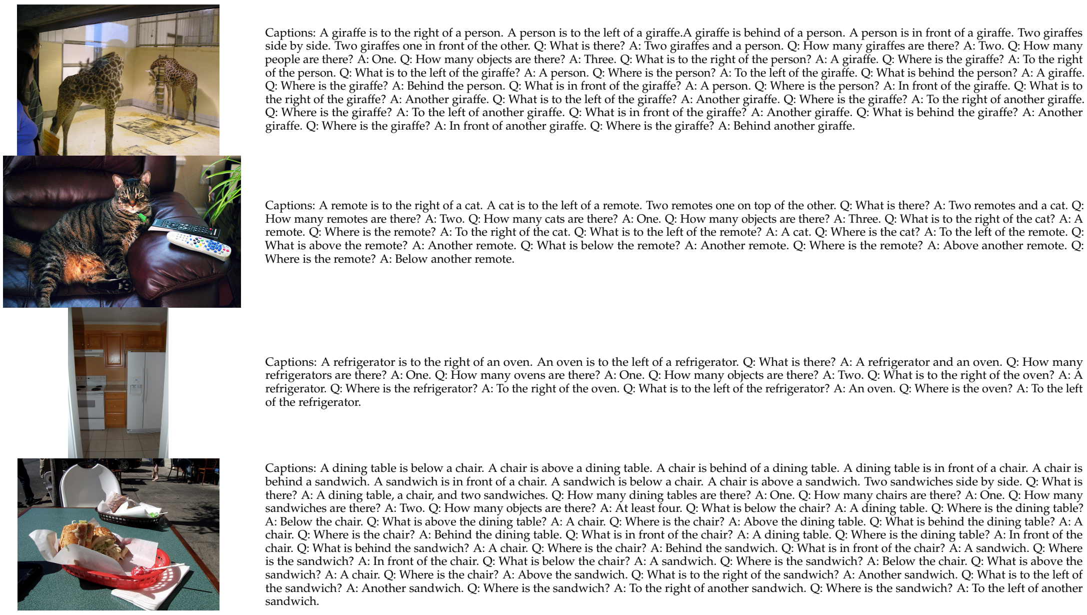


Figure 4. Each original image is accompanied by auxiliary visual modalities: (**A**) original images; (**B**) depth images; (**C**–**F**) class binary masks.

Table 1. SADAMB data splits and their number of elements. Note that a total of 25,551 images from the respective original COCO dataset splits were discarded because they did not contain any recognizable objects.

Data Split	Original Images	Spatial Captions	Spatial Q&A
train	72,784	314,159	989,350
validation	35,359	149,487	472,004
test	71,027	294,422	933,211
Total	179,170	758,068	2,394,565



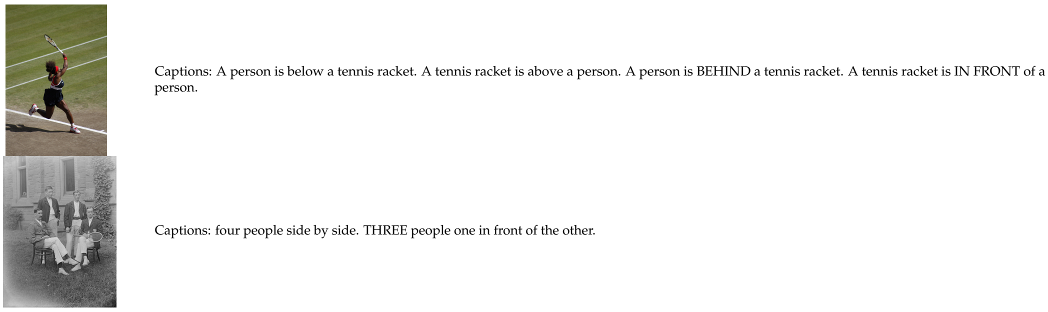
Captions: A giraffe is to the right of a person. A person is to the left of a giraffe.A giraffe is behind of a person. A person is in front of a giraffe. Two giraffes side by side. Two giraffes one in front of the other. Q: What is there? A: Two giraffes and a person. Q: How many giraffes are there? A: Two. Q: How many people are there? A: One. Q: How many objects are there? A: Three. Q: What is to the right of the person? A: A giraffe. Q: Where is the giraffe? A: To the right of the person. Q: What is to the left of the giraffe? A: A person. Q: Where is the person? A: To the left of the giraffe. Q: What is behind the person? A: A giraffe. Q: Where is the giraffe? A: Behind the person. Q: What is in front of the giraffe? A: A person. Q: Where is the person? A: In front of the giraffe. Q: What is to the right of the giraffe? A: Another giraffe. Q: What is to the left of the giraffe? A: Another giraffe. Q: Where is the giraffe? A: To the right of another giraffe. Q: Where is the giraffe? A: To the left of another giraffe. Q: What is in front of the giraffe? A: Another giraffe. Q: What is behind the giraffe? A: Another giraffe. Q: Where is the giraffe? A: In front of another giraffe. Q: Where is the giraffe? A: Behind another giraffe.

Captions: A remote is to the right of a cat. A cat is to the left of a remote. Two remotes one on top of the other. Q: What is there? A: Two remotes and a cat. Q: How many remotes are there? A: Two. Q: How many cats are there? A: One. Q: How many objects are there? A: Three. Q: What is to the right of the cat? A: A remote. Q: Where is the remote? A: To the right of the cat. Q: What is to the left of the remote? A: A cat. Q: Where is the cat? A: To the left of the remote. Q: What is above the remote? A: Another remote. Q: What is below the remote? A: Another remote. Q: Where is the remote? A: Above another remote. Q: Where is the remote? A: Below another remote.

Captions: A refrigerator is to the right of an oven. An oven is to the left of a refrigerator. Q: What is there? A: A refrigerator and an oven. Q: How many refrigerators are there? A: One. Q: How many ovens are there? A: One. Q: How many objects are there? A: Two. Q: What is to the right of the oven? A: A refrigerator. Q: Where is the refrigerator? A: To the right of the oven. Q: What is to the left of the refrigerator? A: An oven. Q: Where is the oven? A: To the left of the refrigerator.

Captions: A dining table is below a chair. A chair is above a dining table. A chair is behind of a dining table. A dining table is in front of a chair. A chair is behind a sandwich. A sandwich is in front of a chair. A sandwich is below a chair. A chair is above a sandwich. Two sandwiches side by side. Q: What is there? A: A dining table, a chair, and two sandwiches. Q: How many dining tables are there? A: One. Q: How many chairs are there? A: One. Q: How many sandwiches are there? A: Two. Q: How many objects are there? A: At least four. Q: What is below the chair? A: A dining table. Q: Where is the dining table? A: Below the chair. Q: What is above the dining table? A: A chair. Q: Where is the chair? A: Above the dining table. Q: What is behind the dining table? A: A chair. Q: Where is the chair? A: Behind the dining table. Q: What is in front of the chair? A: A dining table. Q: Where is the dining table? A: In front of the chair. Q: What is behind the sandwich? A: A chair. Q: Where is the chair? A: Behind the sandwich. Q: What is in front of the chair? A: A sandwich. Q: Where is the sandwich? A: In front of the chair. Q: What is below the chair? A: A sandwich. Q: Where is the sandwich? A: Below the chair. Q: What is above the sandwich? A: A chair. Q: Where is the chair? A: Above the sandwich. Q: What is to the right of the sandwich? A: Another sandwich. Q: What is to the left of the sandwich? A: Another sandwich. Q: Where is the sandwich? A: To the right of another sandwich. Q: Where is the sandwich? A: To the left of another sandwich.

Figure 5. Samples from the SADAMB dataset. For each image, a set of captions and related question–answer pairs are provided.



Captions: A person is below a tennis racket. A tennis racket is above a person. A person is BEHIND a tennis racket. A tennis racket is IN FRONT of a person.

Captions: four people side by side. THREE people one in front of the other.

Figure 6. Erroneous dataset samples stemming from occasional inaccurate depth estimation (relative to the apparent order of the objects) and innate heuristic rule limitations. Errors are presented in UPPERCASE characters.

4. Evaluation Metrics

To define and formulate the metric for evaluating our image captioning models with respect to spatial relations, we will start by breaking down the details of our proposed metric and then provide its theoretical justification and properties.

4.1. Definitions and Formulation

By setting the token size to 500, the models tend to generate more than one sentence per image, which we segment up to the last complete sentence and use for evaluation. The evaluation of spatial relationships is performed through exact matching between the predicted sentences and the ground truth captions of the corresponding image, ensuring a strict and consistent comparison criterion.

A natural question that arises is, "why is exact matching between a generated caption and the corresponding image's ground truth considered a reliable and fair method for evaluating correctness?" Through empirical observations, we found that in our specific use case, powerful models are trained on a simplified vocabulary composed mainly of object labels (e.g., "the cat", "the car", "the person") and a small set of phrases indicating spatial relations (e.g., "is in front of", "is below", "is to the right of"). Within this limited linguistic scope, the models, after only a couple of training epochs, rarely fail to produce grammatically valid sentences. Errors arise almost exclusively from object misidentification or incorrect spatial positioning, precisely the aspects we aim to evaluate. Having said that, the next important question that may arise is, "then, why do we not just use BLEU or any other standard metric instead of the custom one?" The answer to this reasonable inquiry is that, in the context of our intention to focus on the identified objects and their spatial relations, a slight variation in the caption might be completely wrong, even if in BLEU semantics it can be considered of acceptable level (e.g. when the predicted sentence is similar to the reference one, with their only difference being "left" in place of "right"). In this case, BLEU, and any standard metric, would still yield a high score, given that linguistically the two sentences have a large overlap. Thus, the exact match of the generated captions with the ground truth sentences of the image, in the very specific context of our benchmark, is considered to be both reliable and fair. The same reasoning lies also in our choice to not include even more advanced conventional metrics in our comparison (i.e. BERTScore, which compensates for generative synonym creation), since their innate benefits would not manifest and significantly alter the results, and more importantly, the performance order, of the setups tested in this work.

Lastly, a predicted sentence is considered correct if it appears, regardless of its order, anywhere within the ground truth paragraph of the image. We should note that, before matching, a pre-processing step is applied to both ground truth and predicted descriptions where redundant and trailing spaces are stripped and the text is lower-case to ensure consistency and fairness in the comparison.

Given that the task focuses on generating captions that describe spatial relations between objects in images, our evaluation metric consists of five different values, based on the produced sentences, as follows:

Single-Sentence Accuracy: The accuracy of a single caption produced by the model compared to the ground truth caption for a given image, considering spatial relations described in one sentence.

Two-Sentence Accuracy: The accuracy of the first two sentences produced by the model, measuring their alignment with the ground truth description.

Three-Sentence Accuracy: The accuracy of the first three sentences, extending the evaluation to multiple sentence outputs.

All-Sentence Accuracy: The total number of sentences produced across the entire dataset (for all images), comparing them against the corresponding ground truth captions for spatial relations.

Image-Level Accuracy: This metric evaluates how well each image's caption, generated by the model, matches the ground truth, considering spatial relations and object placements.

4.2. Formulaic Representation

In alignment with the prevailing consensus, we acknowledge that a single metric cannot encapsulate all facets of performance in this complex task. Therefore, we introduce a parametrizable metric designed to address two pivotal questions:

- **A**: What is the probability that a generated caption accurately describes a valid spatial relationship?
- **B**: What is the probability of an image being correctly captioned, i.e., to have at least one caption correctly describing a spatial relationship in it?

Our models generate multiple sentences per image, forming a paragraph. Consequently, multiple captions may correspond to a single image, and the number of generated captions can vary across images. We define y as the number of sentences considered per image, with $y \in \{1, 2, 3, \text{max}\}$, where `max` represents the maximum number of captions generated for an image.

4.2.1. Sentence-Level Accuracy (Acc_{yA})

This metric evaluates the proportion of generated captions that exactly match the ground truth captions for their respective images

$$\text{Acc}_{\text{yA}} = \frac{n_{\text{True}}}{n_{\text{True}} + n_{\text{False}}} \tag{1}$$

where

- n_{True} is the total number of generated captions that exactly match any ground truth caption for their corresponding images.
- n_{False} is the total number of generated captions that do not match any ground truth caption.

In the case where $y = 1$, each image contributes one caption, so $n_{\text{True}} + n_{\text{False}} = N$, the total number of images.

4.2.2. Image-Level Accuracy (Acc_{yB})

This metric assesses the proportion of images for which at least one generated caption exactly matches a ground truth caption

$$\text{Acc}_{\text{maxB}} = \frac{1}{N} \sum_{i=1}^{N} \frac{n_{\text{True}}^{(i)}}{\text{max}_i} \tag{2}$$

where

- N is the total number of images;
- $n_{\text{True}}^{(i)}$ is the number of generated captions for image i that exactly match any ground truth caption;
- max_i is the total number of captions generated for image i.

Note that for $y \in \{1, 2, 3\}$, the image-level accuracy Acc_{yB} is equivalent to the sentence-level accuracy Acc_{yA}, as each image contributes the same number of captions.

This formulation provides a comprehensive framework for evaluating model performance in generating spatially accurate captions, accommodating variability in the number of captions per image.

4.3. Properties of the Metric

Granularity: The metric evaluates accuracy at both the sentence and image level, ensuring that fine-grained performance across various levels of output (from individual sentences to full image captions) is captured.

Spatial Context Awareness: It is specifically designed to capture the spatial relationships between objects in the image, which is the core aspect of our task. This makes the metric more relevant and targeted for spatial captioning tasks compared to traditional image captioning metrics.

Flexibility: By evaluating the accuracy of the first few sentences, as well as the total number of sentences generated, the metric can give insights into both the precision of initial predictions and the ability of the model to generate comprehensive and consistent captions.

Sentence-Level Accuracy: The proposed metric goes beyond overall accuracy and measures the correctness of individual sentences, which is crucial in multi-sentence captioning tasks where each sentence might describe different aspects of spatial relationships.

Image-Level Evaluation: By including an image-level accuracy measure, the metric assesses how well the entire set of generated sentences matches the overall ground truth description for each image, enabling a different evaluation perspective that might be more suitable to some uses.

5. Benchmark and Results

5.1. Description of the Models Tested

For our benchmarking, we utilized a combination of popular Hugging Face [25] pre-trained models for the vision and language modalities. Specifically, we fine-tuned the pre-trained versions of ViT (Vision Transformer) (vit-base-patch16-224 from the google/vit-base-patch16-224 repository), DeiT (Data-efficient Image Transformer) (deit-base-patch16-224 from the facebook/deit-base-patch16-224 repository), and DPT (Dense Prediction Transformer) (dpt-large from the Intel/dpt-large repository) as vision encoders, paired with GPT-2 (gpt2 from the gpt2 repository) and BERT (bert-base-uncased from the bert-base-uncased repository) as text decoders, the latter with special provisions. We should note here that we chose to focus on powerful, yet small, edge-device-ready architectures, on a limited number of encoder–decoder combinations that would allow for reasonable experiment durations. We consider the main contribution of our work to be the spatial extension of the COCO dataset, while the benchmark study serves the supplementary purpose of demonstrating a first use of the dataset, utilizing somewhat basic architectures in an interesting combinatorial context. Below is a description of each model and how they were employed in the context of this work.

5.1.1. Vision Encoders

ViT (Vision Transformer): Dosovitskiy et al. [9] treats images as sequences of fixed-size patches, which are then fed into a transformer model to capture long-range dependencies within the image. We fine-tuned it on our spatially enriched dataset to leverage its ability to model spatial relationships between objects in images. Being synonymous with the state of the art, the ViT encoder is easily the go-to option for almost any vision task nowadays.

DeiT (Data-Efficient Image Transformer): Touvron et al. [26] is an efficient variant of the ViT model, designed to require fewer training resources without sacrificing performance. By incorporating distillation from a convolutional neural network (CNN), DeiT achieves competitive results with significantly reduced computational cost. We selected DeiT for its efficiency in fine-tuning on our spatially enriched dataset.

DPT (Dense Prediction Transformer): Ranftl et al. [27] is another variant of the Vision Transformer, optimized for dense prediction tasks such as depth estimation and segmentation. Although DPT was originally designed for dense pixel-level predictions, we tested it here on a high-level spatial reasoning task in the context of IC. The ability of DPT to capture fine-grained spatial relationships and context within images made it a valuable addition to the set of encoders used in this benchmark.

5.1.2. Text Decoders

GPT-2 (Generative Pretrained Transformer 2): Radford et al. [10] is a large-scale transformer model designed for autoregressive text generation. The implementation we used has been pre-trained on large amounts of text data and is capable of generating coherent and contextually appropriate text based on a given prompt. We used GPT-2 as one of the two options for the text decoder to generate captions from the image features extracted by the ViT, DeiT, or DPT encoders, respectively.

BERT (Bidirectional Encoder Representations from Transformers): Devlin et al. [28] is a bidirectional transformer model designed for a variety of language understanding tasks. Unlike GPT-2, BERT is trained to understand context in both directions (left-to-right and right-to-left), which allows it to perform exceptionally well on tasks such as question answering, sentiment analysis, and text classification. BERT's ability to handle complex, bidirectional dependencies made it a strong candidate for tasks requiring a nuanced understanding of spatial relationships between objects. In its primal form, BERT is an encoder architecture. Although not common in the literature, BERT in the Hugging Face framework can be adapted to behave as a decoder. This involves the addition of a language-modeling head, tweaking the model manually to attend to encoder outputs, implement the cross-attention scheme, and define a generation-boundaries workaround that involves the re-purposing of 'PAD' tokens for all 'EOS', 'CLS', and 'PAD' functionalities.

5.2. Fine-Tuning

For all encoder–decoder combinations, we fine-tuned the produced models in an end-to-end manner. This enabled the models to adapt their feature extraction capabilities to the specific requirements of generating spatially aware captions or answering spatiality-related questions. We used the AdamW [29] optimizer with a learning rate of 5×10^{-5} and weight decay of 1×10^{-4} across all experimental instances, and employed cosine annealing learning rate scheduling to promote fast convergence. Cross-entropy loss was used for caption generation. To streamline our training pipeline and manage configurations more elegantly, we employed the Hydra [30] framework and we used MLflow to log and visualize training metrics throughout the fine-tuning process.

5.3. Benchmarking and Experimental Setup

For our purposes, we evaluated the performance of each encoder–decoder combination, assessing how well each model can generate accurate and detailed captions that recognize the objects in the images and describe correctly the spatial relationships between them, as captured by our chosen metrics. We should note here that for the auto-regressive caption generation we used greed-decoding, performing no sampling. In our experiments, all models were trained for 50 epochs using a batch size of 8 and an image resolution of 224. We used a maximum token length of 500 and employed an iterable dataset setup. All training was conducted on two NVIDIA RTX 3090 GPUs, each with 24 GB of VRAM, using data parallelism. The system was powered by an Intel(R) Core(TM) i9-10900X CPU @ 3.70GHz with 64 GB of RAM.

5.4. Results and Analysis

Table 2 summarizes the performance of various vision–text transformer combinations evaluated on standard and proposed spatially aware metrics. Across all metrics, the DeiT/BERT model consistently outperforms other architectures, achieving the highest scores in Bleu (70.69), Rouge (79.62), and Meteor (78.06), indicating its superior linguistic and semantic alignment capabilities, which is a surprising and useful result given that this specific combination is not adequately highlighted in the existing literature.

When focusing on spatially grounded evaluation, DeiT/BERT also leads across all custom metrics ($\mathrm{Acc}_{1\mathrm{A}}$ to $\mathrm{Acc}_{\mathrm{maxB}}$), suggesting that this combination is more adept at capturing and articulating spatial relations within captions. Notably, it achieves a substantial lead in $\mathrm{Acc}_{3\mathrm{A}}$ (45.21) and $\mathrm{Acc}_{\mathrm{maxA}}$ (41.44), which measure deeper, sentence-level spatial accuracy and overall caption-level correctness, respectively.

ViT/BERT follows closely in performance, especially in $\mathrm{Acc}_{1\mathrm{A}}$ (51.54) and $\mathrm{Acc}_{2\mathrm{A}}$ (45.87), while DPT-based models generally underperform compared to ViT and DeiT variants, indicating potential limitations in DPT's spatial generalization capabilities for this task. Additionally, GPT-2 decoders tend to lag behind BERT-based ones across most metrics, reaffirming the benefits of BERT's bidirectional encoding in capturing spatial and contextual dependencies within image-grounded descriptions, an aspect also not broadly mentioned in the literature.

Table 2. Benchmark results for different vision encoder–text decoder combinations fine-tuned for 50 epochs. Best values are denoted in **bold**.

	Performance Scores					
	ViT/GPT2	**ViT/BERT**	**DeiT/GPT2**	**DeiT/BERT**	**DPT/GPT2**	**DPT/BERT**
Bleu	68.9205	69.1500	69.5859	**70.6920**	69.3144	69.1059
Rouge	78.7030	78.8027	78.7846	**79.6154**	78.5726	78.2929
Meteor	77.0582	77.1043	77.2755	**78.0631**	76.8082	76.6636
$\text{Acc}_{1\text{A}}$	51.4664	**51.5399**	50.6547	51.4155	47.7672	48.2565
$\text{Acc}_{2\text{A}}$	44.8609	45.8751	45.6122	**46.7670**	43.9423	43.6683
$\text{Acc}_{3\text{A}}$	42.1286	43.0079	43.8455	**45.2108**	41.9176	42.6292
Acc_{maxA}	38.6304	38.4833	39.8153	**41.4448**	39.2559	39.3343
Acc_{maxB}	46.7449	46.6486	46.4213	**47.4918**	44.8170	44.8145

Overall, the results highlight the critical role of both the vision encoder and text decoder in achieving high spatial reasoning accuracy, with DeiT/BERT emerging as the most effective architecture under the spatially aware captioning benchmark.

6. Metrics Observations

We conducted a study to investigate how the metrics used in our benchmark correlate with each other, drawing important insight as seen in Figure 7. We should note that, all reported values are considered statistically significant with $p < 0.01$.

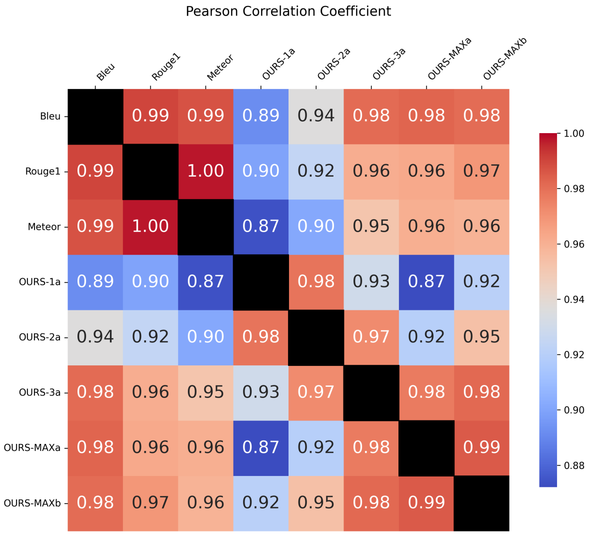


Figure 7. Pearson correlation between the standard and proposed evaluation metrics.

- A value of **1.0** indicates perfect positive correlation.
- A value of **0.0** indicates no correlation.
- A value of **<0** would indicate negative correlation (none observed here).

6.1. Key Observations

1. **Standard metrics are tightly correlated.**
   Metrics like `Bleu` and `Rouge1` correlate strongly ($\geq$0.99), indicating
   - High agreement on the quality of the caption.
   - Redundancy—using all may provide limited additional insight.

2. **Custom spatial metrics correlate strongly among themselves.** Examples include
   - $\texttt{Acc}_{1A}$ vs. $\texttt{Acc}_{2A}$: 0.978;
   - $\texttt{Acc}_{3A}$ vs. $\texttt{Acc}_{\text{maxA}}$: 0.976.

   This suggests coherence across custom metrics targeting different prediction depths.
3. **There are variations of the custom vs. standard metrics which exhibit a correlation gap.**
   - `Meteor` vs. $\texttt{Acc}_{1A}$: 0.87;
   - `Meteor` vs. $\texttt{Acc}_{2A}$: 0.90;
   - `Meteor` vs. $\texttt{Acc}_{3A}$: 0.95.

   This indicates that there are varying levels of alignment depending on the targeted prediction depth. $\texttt{Acc}_{1A}$ and $\texttt{Acc}_{2A}$ seem the most valuable, providing insight possibly lost if only standard metrics were to be used.
4. **MAX metrics are robust aggregators.**
   $\texttt{Acc}_{\text{maxA}}$ and $\texttt{Acc}_{\text{maxB}}$ show a very high correlation (0.985), implying near interchangeability. They also correlate consistently with other metrics, reflecting their role as composite indicators.

6.2. Implications

The correlation analysis suggests that our metric family **SCA** provides a valid, stable, and complementary evaluation signal for spatial image captioning. While traditional metrics remain useful for assessing language fluency, they are insufficient for evaluating spatial relationships. In contrast, **SCA** directly targets the core challenge of our task, accurately capturing spatial relations between objects. Therefore, we recommend using both types of metrics (especially the $\texttt{Acc}_{1A}$ and $\texttt{Acc}_{2A}$ SCA variants, together with any of the conventional metrics) to holistically evaluate spatial captioning models.

7. Conclusions

In this work we introduced **SADAMB**, a novel dataset designed to advance spatially aware image captioning and visual question answering. To introduce our dataset impactfully, we conducted a relatively unexplored vision encoder–text decoder combination benchmark to discover and propose the most efficient architecture suitable for spatial image captioning tasks. In this way, we also propose a new family of metrics, **SCA**, tailored to evaluate spatial understanding in image captioning, providing deeper insights into the models' capabilities. Together, all these contributions aim to propel research in spatially aware image captioning and visual question answering by offering robust data, knowledge, and evaluation methodologies.

Author Contributions: Conceptualization, P.D., G.P., and A.D.; methodology, A.D. and P.D.; software, G.P. and A.N.; validation, A.D. and P.D.; formal analysis, G.P.; investigation, G.P.; resources, D.Z.; data curation, G.P. and A.N.; writing—original draft preparation, G.P.; writing—review and editing, P.D. and A.D.; visualization, G.P.; supervision, D.Z.; project administration, P.D.; funding acquisition, D.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the EU Project VOXReality (Voice-driven Interaction in XR Spaces) under Grant 101070521. The views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or Directorate-General for Communications Networks, Content and Technology. Neither the European Union nor the granting authority can be held responsible for them. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Data Availability Statement: The original data presented in this study are openly available in Zenodo at https://zenodo.org/records/16420621?preview=1&token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6ImMwZGQ2M2EwLTFhYmUtNDVhOS05ODVmLWE1ODQ2MGRmOGJlZiIsImRhdGEiOnt9LCJyYW5kb20iOiI1MTI5ZTQxYTcyMjY4YmYwNjk1YWI3OTljYjU2ZjRhOSJ9.jwZqZlrFKvZPZXXEjyizsr-WJCMaCf6SPXDCVlvUnv6901s9DkCXrQI8Jv1UoZW_XBrxPVlmck0Vz8Ofm0llWg. The raw data supporting the conclusions of this article will be made available by the authors on request.

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

References

1. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
2. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXX 16; Springer: Cham, Switzerland, 2020; pp. 121–137.
3. Lu, J.; Batra, D.; Parikh, D.; Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 13–23.
4. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
5. Bojarski, M.; Del Testa, D.; Dworakowski, D.; Firner, B.; Flepp, B.; Goyal, P.; Jackel, L.D.; Monfort, M.; Muller, U.; Zhang, J.; et al. End to end learning for self-driving cars. *arXiv* **2016**, arXiv:1604.07316. [CrossRef]
6. Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I.; Gould, S.; Van Den Hengel, A. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3674–3683.
7. Gu, S.; Holly, E.; Lillicrap, T.; Levine, S. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 3389–3396.
8. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J.A.; Van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
9. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
10. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
11. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
12. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, 29 June 2005; pp. 65–72.
13. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 74–81.
14. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; proceedings, part v 13; Springer: Cham, Switzerland, 2014; pp. 740–755.
15. 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, Santiago, Chile, 7–13 December 2015; pp. 2425–2433.
16. Liu, R.; Liu, C.; Bai, Y.; Yuille, A.L. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4185–4194.
17. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 3104–3112.
18. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.J.; Shamma, D.A.; et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis. (IJCV)* **2017**, *123*, 32–73. [CrossRef]
19. Johnson, J.; Hariharan, B.; Van Der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2901–2910.

20. Hudson, D.A.; Manning, C.D. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. *arXiv* **2019**, arXiv:1902.09506. [CrossRef]
21. Storks, S.; Gao, Q.; Chai, J.Y. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *arXiv* **2019**, 1–60, arXiv:1904.01172.
22. Liu, F.; Emerson, G.; Collier, N. Visual Spatial Reasoning. *Trans. Assoc. Comput. Linguist.* **2023**, *11*, 635–651. [CrossRef]
23. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
24. Bhat, S.F.; Birkl, R.; Wofk, D.; Wonka, P.; Müller, M. ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth. *arXiv* **2023**, arXiv:2302.12288.
25. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45.
26. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, PMLR, Online, 18–24 July 2021; pp. 10347–10357.
27. Ranftl, R.; Bochkovskiy, A.; Koltun, V. Vision transformers for dense prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 12179–12188.
28. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4171–4186.
29. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
30. Yadan, O. Hydra—A Framework for Elegantly Configuring Complex Applications. Github. 2019. Available online: https://hydra.cc/ (accessed on 17 September 2025).

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.