IJICTR
International Journal of Information & Communication Technology Research

Volume 17- Number 4 – 2025 (40-49)

Context-aware Image Understanding in VQA with Dense-captioning

Elham Alighardash
RIV Lab., Dept. of Computer Engineering, Faculty of Engineering, Bu-Ali Sina University
Hamedan, Iran
e.alighardash@eng.basu.ac.ir,

Hassan Khotanlou*
RIV Lab., Dept. of Computer Engineering, Faculty of Engineering, Bu-Ali Sina University
Hamedan, Iran
khotanlou@basu.ac.ir

Simon Dobnik
CLASP, Department of Philosophy, Linguistics and Theory of Science (FLoV)
University of Gothenburg
Gothenburg, Sweden
simon.dobnik@gu.se

Received: 12 September 2024 – Revised: 3 August 2024 - Accepted: 22 December 2024

Abstract—**Visual Question Answering (VQA) is a complex task that requires models to jointly analyze visual and textual inputs to generate accurate answers. Reasoning and inference are critical for addressing questions that involve relationships, spatial arrangements, and contextual details within an image. In this study, we propose a model based on the BLIP framework, as a generative model, that enhances contextual understanding by incorporating dense-captions -detailed textual descriptions generated for specific regions within an image- along with spatial information extracted from the image. The model focuses on emphasizing visual information and extracting additional context to improve answer accuracy. Experimental results on the GQA dataset demonstrate that the proposed approach achieves competitive performance compared to state-of-the-art methods.**

Keywords: visual question answering, dense-captioning, BLIP, context-aware VQA, reasoning

Article type: Research Article

I. INTRODUCTION

Enhancing the model's understanding of the visual content and strengthening the integration of visual information—considering its greater complexity compared to textual information in the Visual Question Answering (VQA) task—is a promising approach for improving answer prediction accuracy. VQA models often ignore visual information due to the simplicity and ease of learning from language signals [1] or, at the very least, fail to pay sufficient attention to all the necessary details when answering to textual questions about images. However, incorporating contextual information can potentially enhance the model's ability to comprehend the scene more deeply.

Contextual information extends beyond the mere identification of objects and their attributes within an image and encompasses both visual and non-visual information related to the image modality. This information provides models with critical cues to resolve ambiguities and establish connections between concepts. Contextual information can be categorized into appearance context, such as the colors or shapes present in an object's background, and semantic context, which reflects the likelihood of certain objects or events occurring in specific scenes over others [2]. Semantic context can further be divided into spatial semantics, which relates to the arrangement of objects

* Corresponding Author

within a scene, and temporal semantics, which pertains to the timing and sequence of events.

Dense captioning is a computer vision task designed to generate multiple captions for different regions or objects within an image, offering detailed and localized descriptions [3]. In Fig. 1, several dense captions are shown for their corresponding images. Basically, this task involves multiple steps, including feature extraction, region proposal, and region caption generation.

In dense-captioning, several types of contextual information are extracted from an image, like object-level, spatial, semantic, interaction, activity, and temporal contexts. Object-level context identifies individual objects and their attributes, while spatial context clarifies the positioning of various scene components relative to each other. Semantic context captures the overall composition of the scene by describing how different objects and elements contribute to the broader context. Interactions between objects or between people and describing detail activities or actions being performed by people within the image is also considered in dense-captioning.

The contextual information generated by dense captioning can support tasks like VQA by providing detailed background information that enhances the model's ability to answer questions about the scene. Furthermore, translating images to text through dense captioning enhances the extraction and utilization of visual information by providing additional semantic information [4].

In this study, we investigated the impact of incorporating dense image captions on the quality of answers in the VQA task. To achieve this, we utilized the base BLIP [5] model and examined the effects of integrating captions generated by the GRiT [6] model. The study was conducted in two stages, focusing on questions related to relationships between entities in the GQA dataset [7]. These questions were selected due to their complexity and the model's lower performance in answering them, as observed in our previous research.

The main contributions of this study are as follows:

- We propose an approach that incorporates dense-captions generated by a generative model to enrich contextual information, enhancing the performance of Visual Question Answering (VQA)systems.
- We emphasize evaluating the model’s performance on complex relationship-based questions from the GQA dataset, addressing an area where existing models often underperform.
- A two-stage experimental setup is designed to systematically analyze the impact of dense captions, providing insights into their role in improving relational reasoning.

The rest of this paper is organized as follows: Section 2 reviews related work on image dense-captioning and the use of contextual information. Section 3 discusses VQA models and their evolution. Section 4 examines previous studies that have attempted to merge these two tasks. Section 5 outlines the proposed methodology, including the integration of dense captions and the experimental setup. Section 6 presents the results and provides a discussion of the findings. Finally, Section 7 concludes the paper and suggests directions for future research.

II. Image Dense-Captioning Approaches

Early dense-captioning models focused on identifying object-containing regions and describing them, typically based on object detection models such as Faster R-CNN [8]. However, challenges such as disregarding the logical relationship between the features of each region and the overall image features, as well as the aperture problem [9], affected the final description.

In order to address these issues, other models and methods were developed following the primary approaches. The first different perspective, which considered context, was introduced by [10] which merge global and local information to generate a caption for a specific ROI. Additionally, in [11], information from neighboring regions was used to construct a similarity graph, enhancing the representation of relationships in the description of each region. Other research [12] has also attempted to achieve better results for this task by placing greater emphasis on objects within the image as important informational cues.

With the introduction of transformers and visual and language pre-trained (VLP) models, the use of this architecture has also gained attention in dense-captioning. Transformer-based Dense Captioner (TDC) [13] implemented a Region-Object Correlation Scoring Unit (ROCSU) to highlight prominent image regions, treating them as more informative visual content. The CapDet method [14] outperformed previously proposed models for the open-vocabulary object detection task. It is capable of detecting objects from a predefined category list and generating description for newly seen categories in real-world scenes.

Understanding open-world scenes is primarily based on two tasks: Open World Object Detection (OWD) [15] and Open Vocabulary Object Detection (OVD) [16]. Although both of these tasks are aimed to improve object detection systems' ability to generalize to novel categories but they adopt different approaches. OWD involves continuous learning and handles the challenge of detecting unknown classes and incrementally learning them as they become labeled. While OVD aims to extend object detection to a large number of object categories using semantic information from the vocabulary for detecting unseen objects without additional training.

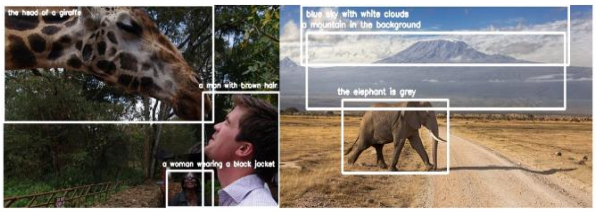

Figure 1. Images and some dense-captions

Generative Region-to-Text Transformer (GRiT) [6] transitions from traditional R-CNN-based architectures to transformer-based approaches, such as Vision Transformer (ViT) [17], which alleviates the need for complex object relationship perception and context modeling. It unifies both object detection and dense captioning by operating in open-set settings and continuously learning newly encountered object classes. The authors believe that, inspired by the idea of generative image-to-text transformers, they have offered a free-form of words and sentences that is more challenging in comparison with open vocabulary and closed-set object detection. However, they achieved comparable results on the COCO dataset [18].

III. VQA Models

Following the introduction of Visual Question Answering (VQA) as a multimodal task combining visual and textual information [19, 20, 21, 22], significant research efforts have focused on this field. Early approaches to VQA explored various model architectures and methods, including efficient object recognition models [23] like ResNet [24] and VGGNet [25] (originally developed for image classification), recurrent models [26, 27, 28], attention mechanisms [29, 30], and memory-augmented networks [31]. Many of these foundational approaches were further refined in subsequent years.

From 2017 onward, there was a shift towards enhancing models' reasoning capabilities [32, 33], employing advanced forms of attention [34,35], and incorporating bi-linear networks [36, 37] and graph-based architectures [38, 39, 40, 41]. The emergence of transformer-based models introduced a new trajectory for VQA research [42, 43, 44]. This line of study evolved further with the integration of pre-training and fine-tuning strategies, utilizing pre-trained vision-language models [5, 45, 46, 47] to improve performance.

The BLIP (Bootstrapping Language-Image Pre-training) model [5], proposed in 2022, is a vision-language framework designed to complete many tasks including VQA. It utilizes a transformer-based architecture that integrates vision encoders, such as ViT [17], with BERT [48] text encoders and decoders to enable effective cross-modal understanding. BLIP employs a multi-stage pre-training strategy, including image-text matching, image-grounded text generation, and contrastive learning, to learn rich and semantically aligned representations. Unlike traditional classification-based approaches for VQA, it predicts answers using a generation-based method, allowing to produce more flexible and contextually accurate responses. Its flexible architecture supports both discriminative and generative tasks, making it highly effective for applications that require reasoning, inference, and contextual understanding of visual and textual data.

Numerous efforts have aimed to enhance these models by incorporating additional contextual information. Some approaches integrate external knowledge sources, such as knowledge graphs [49, 50] and language models [51, 52, 53], along with advanced attention mechanisms to improve the interpretation and alignment of visual and textual components. While this paper does not encompass all related studies, it underscores the critical role of contextual understanding in generating accurate responses.

IV. Dense-captioning in VQA

The parallelism between image captioning and visual question answering, both of which emphasize linking visual and textual aspects of an image, has motivated researchers to explore using image-generated textual descriptions to enhance question answering performance on the same images. Detailed descriptions of various image parts, as well as more abstract, concise captions, can serve as a form of textual translation of the visual content. Leveraging such contextual information helps enrich the visual representation and can mitigate the bias toward textual information—one of the key challenges [1] in VQA tasks.

One of the earliest successful studies [54] to apply the attention mechanism for integrating image caption information with other inputs effectively utilized the LSTM architecture, outperforming previous methods. In this approach, essential captions were automatically generated using data from the VQA v2 dataset [1]. In another study [55] conducted in 2019, a similar objective was pursued. To achieve this, the captioning module functioned as a supplementary knowledge source to enhance VQA performance, offering additional cues for more accurate object identification and refining top-down attention weights. They implemented a two-layer GRU that processed question-attended image features, question embeddings, and caption data to generate final caption features. This approach ensures that captions are effectively filtered, allowing only question-relevant captions to contribute to the VQA process. In fact, the similarity between the captions and the question serves as a crucial criterion for selecting captions during model training.

In knowledge-based Visual Question Answering (KVQA), significant efforts have been directed toward incorporating external knowledge sources to enrich the input information and enhance the model's reasoning capabilities. A substantial challenge has been selecting sufficiently fine-grained and relevant information that aligns with both the visual content and the posed questions. In [56] by constructing a semantic graph from dense-captions of the input image and integrating it with a fact graph and a visual graph—wherein relationships between objects within the image are delineated—the model's reasoning capacity has been notably enhanced. Dense captioning was employed to uncover local-level semantics within the image, capturing details about objects, their attributes, and relationships such as actions, spatial positioning, and comparative relations between objects.

Recent research has re-examined the role of dense-captioning in uncovering contextual layers of visual information. A recent study [57] on VQA explores the integration of large language models (LLMs) with image captioning techniques. The performance of this approach is compared against zero-shot state-of-the-art (SOTA) VQA models such as BLIP-2 [46] and CogVLM [58]. Zero-shot VQA refers to a model's ability to answer questions about images without prior

exposure to similar image-question-answer pairs during training. This method generates two types of captions—general-purpose captions and question-specific captions tailored to the query. To improve the quality of the QA process, the approach selectively retains only the most relevant parts of these captions, reducing the influence of irrelevant or noisy information. Effectively, this method reformulates the VQA task into a question-answering (QA) problem by leveraging dense captions generated using CogVLM or BLIP-2 instead of relying directly on visual features.

The incorporation of dense captions has been a focal point in recent VQA research. This study [4] specifically aimed to mitigate the influence of prior linguistic biases while enhancing the contribution of visual information in the reasoning process. To achieve this, the authors introduced a dense caption-aware VQA model and evaluated its performance across multiple benchmarks, including GQA, GQA-OOD [59], VQA v2, and VQA-CP v2 [60]. The results demonstrated that the proposed model which is named DenseCapBert effectively reduced linguistic bias, thereby improving the model's reliance on visual cues for answering questions.

V. PROPOSED METHOD

This section introduces our proposed approach for addressing context-aware VQA by utilizing dense captioning as complementary information to enhance visual reasoning. The method incorporates position-aware dense captions to improve the detection and understanding of relational information, specifically targeting spatial relationships. In contrast to existing approaches, our method prioritizes the transformation of visual content into textual descriptions, enabling a more effective integration of visual and linguistic information. The core concept involves using dense captions to enrich the model's comprehension of visual context and spatial dynamics. The next subsection provides a detailed explanation of how spatial information and dense captions are merged in the proposed model.

To address the constraints on computational resources while preserving contextual attention and enhancing semantic filtering for the elimination of irrelevant information of captions, a hybrid approach was adopted to achieve an optimal balance between efficiency and effectiveness.

To achieve this objective, the GRiT model was first employed to generate dense-captions that provide detailed descriptions of the image regions associated with the posed question. These captions were then processed using a sentence-BERT similarity [61] metric to evaluate their relevance to the question, retaining only the most relevant ones. Additionally, some other dense-captions corresponding to spatially overlapping regions in the image were preserved to maintain spatial and semantic context.

Next, the filtered captions were transformed into embeddings through a Caption Encoder, following the same method as the question embeddings, to prepare them for further processing. The base BLIP architecture was modified to utilize two parallel co-attention mechanisms that establish meaningful associations between the question and the information derived from both textual and visual modalities.

The first co-attention mechanism captures interactions between the question embeddings and the dense-caption embeddings, enabling textual reasoning. Simultaneously, the second co-attention mechanism models interactions between the question embeddings and visual features extracted from the image, enabling visual reasoning.

The outputs of these two co-attention mechanisms are subsequently merged within a Fusion Block. This block concatenates the outputs and applies a linear projection, ensuring that both textual and visual information are integrated into a unified representation. This fused representation is then passed to the Answer Decoder, which employs causal self-attention and cross-attention layers to generate the final answer to the question. The proposed model structure is shown in Fig. 2.

An initial analysis of the basic model's responses to the question types in the target dataset revealed the necessity of placing greater emphasis on spatial relationships and spatial information within the images. Consequently, this study ensures that the spatial information embedded in the image captions, following the pre-processing stage, is adequately considered. The subsequent section provides a detailed discussion on the integration of spatial information and relationships.

A. *spatial presentation and relation*

Traditional approaches to visual feature extraction typically generate embeddings directly from visual information, leading to the loss of associated spatial information. Moreover, spatial relationships between entities are not effectively captured in these methods. While some studies have utilized attention mechanisms or fully connected layers to incorporate spatial information, this approach draws inspiration from [62] conducted in 2022. It aims to explicitly learn the spatial representation associated with each caption within the image, facilitating a more comprehensive semantic understanding of the scene.

According to the referenced study, various methods exist for incorporating spatial information into the embeddings of region descriptions. However, to maintain simplicity and facilitate model training through the optimization of continuous values, this approach employs the integration of normalized bounding box information corresponding to each caption. To represent the location and size of bounding boxes, a 9-dimensional vector is employed, normalized by the dimensions of the image. This vector encodes the center point of the bounding box, the normalized width and height, and the coordinates of two diagonal corner points of the bounding box. The structure of this vector is shown in Equation 1.

$$\left(\frac{c_x}{W}, \frac{c_y}{H}, \frac{w}{W}, \frac{h}{H}, \frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H}, \frac{wh}{WH}\right) \quad (1)$$

The position embedding is then concatenated with the caption embedding and passed through a linear layer to project the vector to the desired dimensionality.

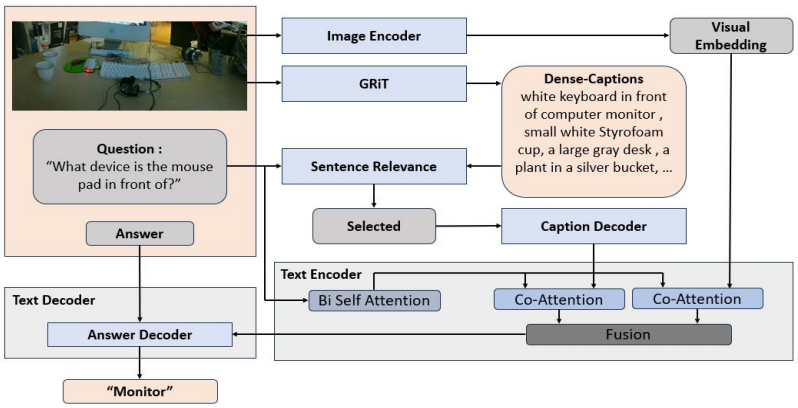

Figure 2. The structure of the proposed model

VI. EVALUATION

A. Experimental setup

In this study, the proposed model was implemented using the BLIP framework as the base architecture, extended with dense captioning and spatial information to enhance contextual understanding. Dense captions were generated for specific regions within each image, providing supplementary textual descriptions to guide the reasoning process. These captions were further enriched with bounding box coordinates to encode explicit spatial relationships.

The model was trained and evaluated on an NVIDIA 3090 GPU with 24 GB of GPU memory, 32 GB RAM, utilizing the learning settings of BLIP in PyTorch. Training was conducted for 15 epochs. The proposed approach was evaluated using the GQA dataset and benchmarked against state-of-the-art VQA models, including BLIP-2 and CogVLM, as well as other caption-augmented VQA methods such as DenseCapBERT, all assessed under comparable experimental conditions.

Performance metrics, including accuracy, METEOR, and BERT-based similarity, were employed to comprehensively evaluate the quality, relevance, and semantic alignment of the predictions.

B. Dataset

We utilize the GQA dataset [7], a large-scale benchmark designed for evaluating visual question answering (VQA) models, based on Visual Genome dataset [63] with a focus on reasoning and compositional abilities. The dataset contains over 22 million questions associated with approximately 113,000 images. For model evaluation, we use the balanced test-dev subset, which includes 12,578 questions pertaining to 398 images. Questions in this subset are categorized based on their structural and semantic characteristics, with each question falling into one of the following groups: object, relation, category, attribute, or global. Previous studies have demonstrated that questions in the relation category are among the most challenging for models to answer, which is why we specifically focus on this category, consisting of 5,038 questions in the test-dev set. Answering these questions requires understanding one or more relationships between objects, which can include spatial, semantic, comparative, property discovery, and other types of relations. The answers generated are tailored to the question type, which may include yes/no (verify), open-ended (query), multiple choice (choose), logical inference (logical), or comparison (compare) types.

C. Results

The evaluation results of the proposed method on relation-based questions from the balanced test-dev subset of the GQA dataset are presented in Table 1. In this table, accuracy—serving as the primary evaluation metric for VQA—is utilized to assess the performance of three models: (1) the BLIP model, (2) the proposed approach leveraging only dense captions (BLIP-Cap), and (3) the final model (BLIP-PCap), which integrates dense captions with positional information derived from bounding box coordinates as input features.

The analysis indicates that augmenting input information with dense captions has led to a general improvement in model performance across the various question types within this category. Moreover, the incorporation of explicit spatial information into dense captions has further enhanced this performance. While the magnitude of improvement varies among different question types, the performance gains exhibit a consistent incremental pattern across the majority of them.

TABLE I. EVALUATION OF DIFFERENT APPROACHES BASED ON ACCURACY- \#TEST-DEV REPRESENTS THE NUMBER OF EACH QUESTION TYPE IN THE RELATION CATEGORY OF THE GQA TEST-DEV BALANCED SET. THIS TABLE COMPARES THE BASELINE PERFORMANCE OF THE BLIP MODEL, THE AUGMENTED MODEL WITH DENSE CAPTIONS (BLIP-CAP), AND THE POSITION-AWARE AUGMENTED BLIP MODEL (BLIP-PCAP).

Relation	#test-dev	BLIP	BLIP-Cap	BLIP-PCap
categoryRelO	339	0.52	0.56	0.6
categoryRelOChoose	41	0.9	0.93	0.94
categoryRelS	929	0.44	0.48	0.49
existRelS	218	0.83	0.87	0.89
existRelSC	148	0.88	0.85	0.87
existRelSRC	28	0.5	0.5	0.51
relChooser	211	0.57	0.51	0.55
relO	1164	0.3	0.35	0.38
relS	1411	0.3	0.36	0.37
relVerify	441	0.86	0.88	0.89
relVerifyCo	273	0.87	0.88	0.9
relVerifyCr	55	0.58	0.56	0.6
sameMaterialRelate	20	0.3	0.35	0.35
sameRelate	30	0.33	0.35	0.35
ALL	5308	0.58	0.6	0.62

It is worth noting that during the caption generation stage, certain image details were overlooked, and in some cases, even the subject of the question was neglected. This limitation, coupled with an insufficient number of similarly structured questions during training, contributed to a reduction in the accuracy of the proposed model. Moreover, such scenarios increased the likelihood of hallucinations in the model. Specifically, for relChooser questions—where the BLIP model frequently struggles to differentiate between left and right orientations—even the inclusion of spatial information in the form of bounding boxes failed to elevate the performance of the proposed model to match that of the original model. However, this limitation was not observed across other question types within this category.

Given that the base model is generative, relying solely on accuracy or exact matching as the evaluation metric was insufficient. Therefore, additional evaluation criteria were employed to provide a more comprehensive assessment of performance. The Fig. 3 illustrates the progression of model improvements through modifications applied during two phases of experiment, evaluated using the METEOR [64] metric.

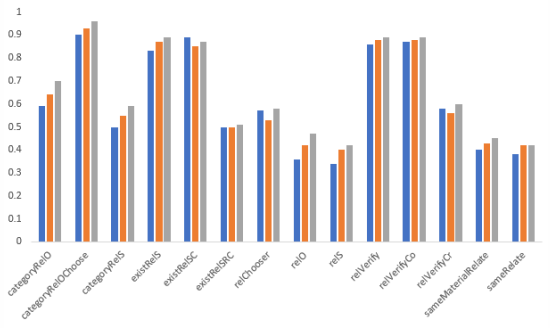

Figure 3. A comparison of different approaches based on Meteor

TABLE II. EVALUATION OF DIFFERENT APPROACHES BASED ON BERT-BASED SIMILARITY

Relation	BLIP	BLIP–Cap	BLIP–PCap
categoryRelO	0.82	0.85	0.88
categoryRelOChoose	0.96	0.99	1
categoryRelS	0.8	0.83	0.84
relChooser	0.93	0.92	0.94
relO	0.75	0.75	0.81
relS	0.78	0.8	0.81
sameMaterialRelate	0.55	0.6	0.6
sameRelate	0.74	0.72	0.72

TABLE III. A COMPARISON OF VQA MODELS BASED ON ACCURACY ON GQA BALANCED TEST-DEV: THE RESULTS OF COGVLM AND BLIP-2 MODELS ARE REPORTED BASED ON ZERO-SHOT VQA SETTINGS, AS DESCRIBED IN [58]

	VQA			VQA + Caption				
Category	LXMERT [42]	BLIP [5]	CogVLM* [58]	BLIP-2* [46]	Ozdemir* [57]	DenseCapBert [4]	BLIP-Cap	BLIP-PCap
Relation	0.64	0.58	0.56	0.40	0.43	-	0.60	0.62
All	0.60	0.56	0.66	0.48	0.49	0.6	0.58	0.61

While traditional evaluation metrics for generative models, such as BLEU [65] and METEOR, are widely recognized and commonly used, they primarily focus on the surface-level lexical similarity of generated outputs, often overlooking their semantic meaning. To address this limitation, the model's performance was additionally evaluated using a BERT-based similarity metric [66]. This approach utilizes contextual embeddings derived from Bidirectional Encoder Representations from Transformers or BERT to capture and assess deeper semantic relationships between the generated outputs and reference texts.

The results of this evaluation are presented in Table 2, offering insights into the semantic alignment of the generated answers. We excluded binary or yes/no questions from this evaluation because such questions inherently have a limited set of answers, and their evaluation using semantic similarity metrics may not accurately reflect performance.

As reported in Table 2, the incorporation of dense captions and spatial information has led to an improvement in the model's contextual understanding. This enhancement is evident from the fact that, in most cases, even when the generated answers are incorrect, they exhibit a high degree of semantic similarity to the reference answers. This observation suggests that the model's overall comprehension of the image and the relationships among its visual elements has improved relative to the base model.

A comparison of the results of the proposed method in two configurations has been conducted against the BLIP model, as a generative approach, and the LXMERT model, as a classification-based approach, which utilize different strategies for answer prediction. The Table 3 also presents a comparative analysis of the performance of CogVLM, recognized as one of the leading VLP frameworks, and the BLIP--2 model, evaluated under zero-shot settings as referenced in [57]. The accuracy results reported for these two models account for responses deemed correct if their cosine similarity with the reference answers exceeded a threshold of 0.7.

Additionally, the performance of the proposed models has been evaluated against other approaches that leverage dense captioning to enhance VQA task. The accuracy of the best-performing method proposed in [58], referred to as Ozdemir, has been evaluated using the same method as CogVLM and blip--2 that are reported in this study. All results have been assessed using accuracy as the evaluation metric, focusing on the relation-based question category within the balanced test-dev subset, as well as the overall questions in this subset. This analysis demonstrates that the proposed model has the potential to compete with state-of-the-art and well-established methods in this domain, while also exhibiting strong performance compared to similar approaches.

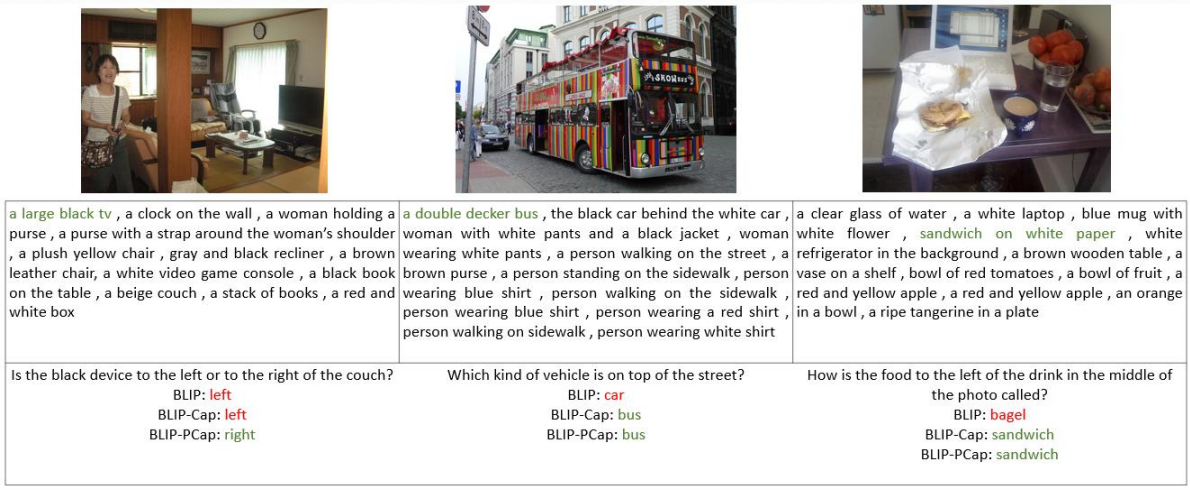

Figure 4. Examples of predictions generated by the baseline BLIP model and the proposed method, including input images (first row), their corresponding dense captions (second row), and associated questions along with the predicted answers

Improvements to the predictions generated by the base model are demonstrated through the proposed approaches, which leverage dense captions and incorporate spatial information from the image in Fig. 4. These examples highlight the effectiveness of generating diverse and contextually relevant captions, combined with a focus on image details, in facilitating the production of accurate answers to the posed questions.

VII. Conclusion

The findings of this study demonstrate that augmenting the initial extracted features in the baseline model with dense caption information enhances background knowledge and improves performance in the VQA task. However, analysis revealed that many of the questions the baseline model failed to answer involved spatial relationships between objects, which were not explicitly captured in the generated captions. To address this limitation, incorporating spatial information derived from the extracted regions during the captioning stage further mitigated deficiencies in the model's performance.

Given the hardware constraints encountered during this study, efforts were directed toward reducing computational costs, which inherently limited the exploration of more complex methods. Future research could benefit from leveraging graph-based structures to extract and integrate spatial information, potentially enabling a more refined understanding of spatial relationships.

Despite the diversity of captions generated by the captioning model, it was observed that these captions occasionally lacked detailed descriptions of all image components and, in some instances, failed to directly reference the entities or attributes mentioned in the questions. This mismatch sometimes led to sub-optimal performance, as the captions provided insufficient contextual support for answering the questions accurately.

To address these challenges, it is recommended that future studies compare the performance of alternative caption generation models in a systematic manner. Additionally, further research could focus on developing captioning approaches that incorporate cues from the given question or emphasize specific spatial relationships between entities mentioned in the question. Such advancements may lead to more effective integration of linguistic and spatial information, further enhancing model performance.

In conclusion, accounting for semantic variations in predicted answers, as well as in captions and questions, can expand the lexical and conceptual space required for comprehending and representing the contextual information inherent in both the image and the question. This broader semantic representation facilitates a more robust understanding of the problem context, thereby improving the model's ability to generate accurate and contextually relevant responses.

References

[1] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter:Elevating the role of image understanding in visual question answering," inProceedings of the IEEEConference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[2] X. Wang and Z. Zhu, "Context understanding in computer vision: A survey," *Computer Vision and Image Understanding*, vol. 229, p. 103646, Mar. 2023, doi: 10.1016/j.cviu.2023.103646.

[3] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," inProceedings of the IEEE Conference on Computer Vision and PatternRecognition (CVPR), June 2016.

[4] Y. Bi, H. Jiang, Y. Hu, Y. Sun, and B. Yin, "See and learn more: Dense caption-aware representation for visual question answering,"IEEE Trans. Cir. and Sys. for Video Technol., vol. 34,no. 2, p. 1135–1146, jul 2023. [Online]. Available: https://doi.org/10.1109/TCSVT.2023.3291379

[5] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," inProceedings of the 39th International Conferenceon Machine Learning, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka,L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162.PMLR, 17–23 Jul 2022, pp.12 888–12 900. [Online]. Available: https://proceedings.mlr.press/v162/li22n.html

[6] J. Wu, J. Wang, Z. Yang, Z. Gan, Z. Liu, J. Yuan, and L. Wang, "GRiT: A generative region-to-text transformer for object understanding," 2022. [Online]. Available: https://arxiv.org/abs/2212.00280

[7] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," inProceedings of the IEEE/CVF Conference on

Computer Visionand Pattern Recognition (CVPR), 2019, pp. 6700–6709.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in Advances in Neural Information Processing Systems (NeurIPS), vol. 28.Curran Associates, Inc., 2015, pp. 91–99.

[9] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," inProceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI),vol. 2, 1981, pp. 674–679.

[10] L. Yang, K. Tang, J. Yang, and L.-J. Li, "Dense captioning with joint inference and visual context,"in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).IEEE, 2017, pp. 2193–2202.

[11] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, and J. Shao, "Context and attribute grounded dense captioning," in2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),2019, pp. 6234–6243.

[12] X. Li, S. Jiang, and J. Han, "Learning object context for dense captioning," in Proceedings of theThirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, ser. AAAI'19/IAAI'19/EAAI'19.AAAI Press, 2019. [Online]. Available:https://doi.org/10.1609/aaai.v33i01.33018650

[13] Z. Shao, J. Han, D. Marnerides, and K. Debattista, "Region-object relation-aware dense captioning via transformer,"IEEE Transactions on Neural Networks and Learning Systems, 2022.

[14] Y. Long, Y. Wen, J. Han, H. Xu, P. Ren, W. Zhang, S. Zhao, and X. Liang, "Capdet: Unifying dense captioning and open-world detection pretraining," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15 233–15 243.

[15] K. J. Joseph, S. Khan, F. S. Khan, and V. N. Balasubramanian, "Towards open world object detection," in2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),2021, pp. 5826–5836.

[16] A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang, "Open-vocabulary object detection using captions," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), June 2021, pp. 14 393–14 402.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani,M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth16x16 words:Transformers for image recognition at scale," 2021. [Online]. Available:https://arxiv.org/abs/2010.11929

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick, "Microsoft coco: Common objects in context," in European Conference on Computer Vision (ECCV).Springer, 2014, pp. 740–755.

[19] W. S. Lasecki, P. Thiha, Y. Zhong, E. Brady, and J. P. Bigham, "Answering visual questions with conversational crowd assistants," in Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS). ACM, 2013, pp. 153–160.

[20] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in Advances in Neural Information Processing Systems 27 (NIPS 2014),Montreal, Quebec, Canada, 2014, pp. 1682–1690.

[21] D. Geman, S. Park, C. Chen, and D. Geman, "Visual turing test for computer vision systems,"Proceedings of the National Academy of Sciences, vol. 112, no. 12, pp. 3618–3623, 2015.

[22] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in Proceedings of the IEEE International Conference on Computer Vision(ICCV), 2015, pp. 2425–2433.

[23] P. Wang, Q. Wu, C. Shen, and A. van den Hengel, "The vqa-machine: Learning how to use existing vision algorithms to answer new questions," in 2017 IEEE Conference on Computer Vision andPattern Recognition (CVPR), 2017, pp. 3909–3918.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedingsof the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition,"arXiv preprint arXiv:1409.1556, 2015.

[26] I. Chowdhury, K. Nguyen, C. Fookes, and S. Sridharan, "A cascaded long short-term memory (lstm) driven generic visual question answering (vqa)," in 2017 IEEE International Conference on Image Processing (ICIP), 2017, pp. 1842–1846.

[27] R. Li and J. Jia, "Visual question answering with question representation update (qru),"in Advances in Neural Information Processing Systems, D. Lee, M. Sugiyama, U. Luxburg,I. Guyon, and R. Garnett, Eds., vol. 29.Curran Associates, Inc., 2016. [Online]. Available:https://proceedings.neurips.cc/paperfiles/paper/2016/file/d69dbe29f156a7ef876a40a94f65599−Paper.pdf

[28] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra, "Visual dialog," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp.326–335.

[29] I. Schwartz, A. Schwing, and T. Hazan, "High-order attention models for visual question answering," Advances in Neural Information Processing Systems, vol. 30, 2017.

[30] Y. Shi, T. Furlanello, S. Zha, and A. Anandkumar, "Question type guided attention in visual question answering," in proceedings of the European conference on computer vision (ECCV), 2018, pp. 151–166.

[31] C. Ma, C. Shen, A. Dick, Q. Wu, P. Wang, A. v. d. Hengel, and I. Reid, "Visual question answering with memory-augmented networks," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6975–6984.

[32] D. A. Hudson and C. D. Manning, "Compositional attention networks for machine reasoning," in International Conference on Learning Representations (ICLR), 2018.

[33] S. R. Venkataraman, R. S. Rao, S. Balasubramanian, R. R. Sarma, and C. S. Vorugunti, "Can you even tell left from right? presenting a new challenge for vqa," in 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4486–4495.

[34] J. Singh, V. Ying, and A. Nutkiewicz, "Attention on attention: Architectures for visual question answering (vqa),"arXiv preprint arXiv:1803.07724, 2018.

[35] H. Sharma and S. Srivastava, "Integrating multimodal features by a two-way co-attention mechanism for visual question answering," Multimedia Tools and Applications, vol. 83, no. 21, pp. 59 577–59 595, 2024.

[36] Z. Yu, J. Yu, J. Fan, and D. Tao, " Multi-modal Factorized Bilinear Pooling with Co-attention Learning for Visual Question Answering ," in 2017 IEEE International Conference on Computer Vision (ICCV).Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2017, pp. 1839–1848. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.202

[37] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in Advances in Neural Information Processing Systems, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paperfiles/paper/2018/file/96ea64f3a1aa2fd00c72faacf0cb8ac9−Paper.pdf

[38] L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for visual question answering," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp.10 313–10 322.

[39] S. Lee, J.-W. Kim, Y. Oh, and J. H. Jeon, "Visual question answering over scene graph," in 2019 First International Conference on Graph Computing (GC), 2019, pp. 45–50.

[40] W. Liang, Y. Jiang, and Z. Liu, "Graghvqa: Language-guided graph neural networks for graph-based visual question answering,"arXiv preprint arXiv:2104.10283, 2021.

[41] Y. Wang, M. Yasunaga, H. Ren, S. Wada, and J. Leskovec, "Vqa-gnn: Reasoning with multimodal knowledge via graph neural networks for visual question

answering," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 21 582–21 592.

[42] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers,"in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), K. Inui, J. Jiang,V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019,pp. 5100–5111. [Online]. Available: https://aclanthology.org/D19-1514

[43] F. Gard`eres, M. Ziaeefard, B. Abeloos, and F. Lecue, "ConceptBert: Concept-aware representation for visual question answering," in Findings of the Association for Computational Linguistics: EMNLP2020, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020,pp. 489–498. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.44

[44] H. Zhong, J. Chen, C. Shen, H. Zhang, J. Huang, and X.-S. Hua, "Self-adaptive neural module transformer for visual question answering," IEEE Transactions on Multimedia, vol. 23, pp. 1264–1273, 2021.

[45] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and vqa," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp.13 041–13 049, Apr. 2020. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/7005

[46] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in European conference on computer vision. Springer, 2020, pp.104–120.

[47] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in Proceedings of the 40th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho,B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 19 730–19 742.[Online]. Available: https://proceedings.mlr.press/v202/li23q.html

[48] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[49] L. Gui, B. Wang, Q. Huang, A. Hauptmann, Y. Bisk, and J. Gao, "Kat: A knowledge augmented transformer for vision-and-language," arXiv preprint arXiv:2112.08614, 2021.

[50] Z. Yang, L. Wu, P. Wen, and P. Chen, "Visual question answering reasoning with external knowledge based on bimodal graph neural network," Electronic Research Archive, vol. 31, no. 4, pp. 1948–1965,2023.

[51] L. H. Li, Y.-C. Su, C. Xiong, S. Sun, and Y. Wang, "Visualbert: A simple and performant baseline for vision and language," arXiv preprint arXiv:1908.03557, 2019.

[52] Z. Hu, P. Yang, Y. Jiang, and Z. Bai, "Prompting large language model with context and pre-answer for knowledge-based vqa," Pattern Recognition, vol. 151, p. 110399, 2024.

[53] J. Lin, H. Yin, W. Ping, P. Molchanov, M. Shoeybi, and S. Han, "Vila: On pre-training for visual language models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2024, pp. 26 689–26 699.

[54] W. Cai and G. Qiu, "Visual question answering algorithm based on image caption," in 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2019,pp. 2076–2079.

[55] J. Wu, Z. Hu, and R. Mooney, "Generating question relevant captions to aid visual question answering," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,A. Korhonen, D. Traum, and L. M`arquez, Eds. Florence, Italy: Association for ComputationalLinguistics, Jul. 2019, pp. 3585–3594. [Online]. Available: https://aclanthology.org/P19-1348

[56] J. Yu, Z. Zhu, Y. Wang, W. Zhang, Y. Hu, and J. Tan, "Cross-modal knowledge reasoning for knowledge-based visual question answering," Pattern Recognition, vol. 108, p. 107563, 2020. [Online].Available: https://www.sciencedirect.com/science/article/pii/S0031320320303666

[57] O. Ozdemir and E. Akag und uz, "Enhancing visual question answering through question-driven image captions as prompts," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2024, pp. 1562–1571.

[58] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Songet al., "Cogvlm: Visual expert for pretrained language models," arXiv preprint arXiv:2311.03079, 2023.

[59] F. Kervadec, G. Antipov, M. Elad, and P. Maragos, "Roses Are Red, Violets Are Blue... But Should VQA Expect Them To?," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8222–8231. DOI: 10.1109/CVPR46437.2021.00814..

[60] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," in Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2018, pp. 4971–4980.

[61] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," arXiv preprint arXiv:1908.10084, 2019.

[62] Y. Duan, Z. Wang, J. Wang, Y.-K. Wang, and C.-T. Lin, "Position-aware image captioning with spatial relation," Neurocomputing, vol. 497, pp. 28–38, 2022.

[63] R. Krishna, S. Antol, J. Hockenmaier, M. Mitchell, and et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," in Proceedings of the IEEE International Conferenceon Computer Vision (ICCV). IEEE, 2017, pp. 1–9.

[64] S. Banerjee and A. Lavie, "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. Association for Computational Linguistics,2005, pp. 65–72.

[65] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). Association for Computational Linguistics, 2002, pp. 311–318.

[66] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in International Conference on Learning Representations, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

Hassan Khotanlou received the B.E. and M.E. degrees in Computer Engineering from Iran University of Science & Technology in 1998 and the Ph.D. degree in Computer Engineering-Machine Vision from Telecom ParisTech in 2007. Since 2007, He has been with the Bu-Ali Sina University and currently He is a Professor of Computer Engineering and Head of the Robot Intelligence and Vision (RIV) Research Group in Bu-Ali Sina University. His current research interests include evolutionary computation, Image and Video Processing, Pattern Recognition and Deep Learning.

Elham Alighardash earned her B.E. in Computer Engineering from Alzahra University and received her M.E. in Artificial Intelligence from Bu-Ali Sina University where she is currently pursuing a Ph.D. in AI at

the RIV Lab. She has been a faculty member at Sayyed Jamaleddin Asadabadi University since 2012. Her research interests include Multi-modal Learning, Digital Image and Video Processing, Natural Language Processing and Machine Learning.

Simon Dobnik received his DPhil (PhD) in Computational Linguistics from the Queen's College, University of Oxford. He is currently a Professor of Computational Linguistics at FLOV and CLASP, University of Gothenburg, Sweden. His research focuses on Computational linguistics, NLP, Spatial Language and cognition.