Enabling Multimodal Understanding: Lidar Data Meets VQA

Muhammad Zeeshan Khan, *Anuroop Gaddam, Dhananjay Thiruvady, and +N.K.Suryadevara

School of Information Technology, Deakin University, Geelong, VIC 3216, Australia,+University of Hyderabad,India
{muhammadz,anuroop.gaddam, dhananjay.thiruvady}@deakin.edu.au

Abstract. This chapter explores the integration of Light Detection and Ranging (LiDAR) data with multimodal systems such as Visual Question Answering (VQA) to enable robust contextual understanding. The chapter begins with an overview of VQA, including its architecture, various modeling approaches, and practical applications. The chapter then introduces LiDAR data, particularly point clouds, and highlights how this 3D information enhances traditional visual input in machine learning tasks. Special attention is given to recent efforts in combining point cloud data with VQA, examining relevant datasets, deep learning models, and fusion techniques. The chapter concludes by outlining current limitations and future directions for advancing LiDAR-based multimodal understanding in VQA.

Keywords: LiDAR, Visual Question Answering (VQA), VQA Applications, Multimodality, Computer Vision, Natural Language Processing

1 Introduction to Visual Question Answering (VQA)

Recent progress in deep learning algorithms has accelerated research in both computer vision (CV) and natural language processing (NLP), along with their combined use in multimodal applications such as Visual Question Answering (VQA). VQA is a field that integrates computer vision and NLP to answer questions about visual content [1]. In VQA systems, computer vision handles the analysis and interpretation of images, while NLP helps to understand and process the associated text or question. Related tasks to VQA in the multimodal domain include image and video captioning [2], as well as retrieval [3]. VQA offers a wide range of real-world applications, such as virtual assistants that help blind users interact with their surroundings [4], chatbots that handle both visual and textual data [5], and even applications in the medical field [6].

In a typical VQA system, the model takes an image and a question as input, and the algorithm predicts an answer in the form of a word or short phrase. There are also variations where answers are provided in binary formats (yes/no), multiple-choice options, or fill-in-the-blank formats. The concept of VQA draws inspiration from Textual Question Answering (TQA), but extends it to handle

N. K. Suryadevara (ed.), *Proceedings of the IoT AND LiDAR Technologies in Healthcare Workshop (ILTH 2024)*, Atlantis Highlights in Intelligent Systems 14,
https://doi.org/10.2991/978-94-6463-784-7_11

complex visual information. VQA is more challenging than related tasks like retrieval or captioning because it requires reasoning and domain-specific knowledge to answer complex questions.

Chapter Scope and Objectives This chapter aims to provide a comprehensive understanding of VQA, its frameworks, and architecture, along with its adaptation to Lidar data, with an emphasis on the integration of Lidar data with multimodalities.

- This chapter provides a structured overview of Visual Question Answering (VQA), covering the foundational frameworks and a variety of model architectures including attention-based methods, reasoning-based models, external knowledge integration, bias mitigation strategies, and recent advances in Vision-Language Models (VLMs). It also highlights the strengths and limitations of each approach in addressing complex real-world queries.
- The chapter provides a detailed overview of LiDAR technology with a focus on the nature of the data it produces—primarily 3D point clouds. It discusses how this data captures geometric and spatial information of the surrounding environment and how it is used in deep learning applications such as 3D object detection, semantic segmentation, and scene understanding within the context of computer vision and multimodal learning.
- A central focus of the chapter is on the intersection of LiDAR data and VQA. It discusses recent efforts in extending VQA to 3D environments through the use of point clouds, and examines benchmark datasets and models that facilitate this task. The chapter emphasizes the importance of multimodal fusion techniques and concludes with future research directions that include integrating LiDAR with large vision-language models for improved generalization and real-world application.

2 Overview of VQA: Framework, Architectures, and Applications

2.1 Framework of VQA Traditionally, in a general Visual Question Answering (VQA) framework, the system takes an input image and a question, then predicts the answer by reasoning over the visual and textual inputs. This process typically follows a sequence of steps: extracting features from both the image and the text, fusing these features, and finally predicting the answer. In contrast, current state-of-the-art Vision-Language Models (VLMs) streamline the entire pipeline by performing feature extraction, fusion, and classification within an end-to-end unified architecture.

In Visual Question Answering (VQA), various deep learning models have been employed for visual feature extraction. These networks are typically designed for specific tasks such as image classification or object detection and are trained on large-scale datasets. For instance, the ImageNet [7] dataset is commonly used for image classification, while COCO [8] and PASCAL VOC [9] datasets are widely used for object detection.

Visual features can be extracted either from the entire image or from local regions identified by object detection algorithms. Examples of classification-based architectures include VGG [10], ResNet [11], GoogLeNet [12], and Vision Transformers [13]. For object detection, models like Faster R-CNN [14] are commonly used.

There are various approaches to encode the textual data used in Visual Question Answering (VQA). The simplest method is one-hot encoding, where each word in the question is converted into a vector representation known as an embedding. Early VQA architectures also utilized basic methods such as Bag-of-Words and Skip-gram models [15].

To capture the sequential nature of textual input, more advanced architectures like Long Short-Term Memory (LSTM) [16] and Gated Recurrent Units (GRU) [17] were introduced. LSTM networks include input, forget, and output gates, while GRUs consist of update and reset gates.

In more recent developments, Bidirectional Encoder Representations from Transformers (BERT) [18], which is composed of several transformer encoder layers, has been widely adopted in VQA systems. Unlike the standard Transformer [19] architecture that includes both encoder and decoder components, BERT uses only the encoder. Most recent VQA architectures have also utilized large language models (LLMs) for text processing.

After extracting visual and textual features in a Visual Question Answering (VQA) system, the next critical phase is feature enhancement. This stage focuses on refining and enriching the extracted features by capturing more meaningful representations within each modality (intra-modality enhancement) and establishing stronger associations between the two modalities (inter-modality alignment). Feature enhancement may also involve integrating external knowledge to supplement reasoning, removing language biases to ensure fair and accurate predictions, and employing compositional modeling to support multi-step and logical reasoning across complex queries. Techniques such as attention mechanisms, graph neural networks, and multimodal transformers are commonly used in this phase.

Following feature enhancement, the next step is feature fusion or concatenation, where the refined visual and textual features are combined into a unified representation. This joint representation is then used for reasoning and answer prediction. The simplest methods used by VQA architectures for feature fusion involve basic operations such as vector concatenation, element-wise addition, and element-wise multiplication between the visual and textual features. While these approaches are straightforward and computationally efficient, they may not fully capture the complex interactions between modalities.

To address this, neural network-based techniques, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have also been employed for feature fusion in VQA. These models allow for more expressive and learnable fusion mechanisms.

In addition, bilinear modeling has gained attention as a powerful technique for feature fusion. This method computes the outer product of visual and textual

features to capture all possible pairwise interactions between elements from both modalities. Although bilinear fusion provides a rich joint representation, it comes with a significant drawback: it is computationally intensive and demands substantial resources, especially when dealing with high-dimensional visual regions and long textual inputs. In recent VQA architectures, transformer-based models have been widely adopted in Vision-Language Models (VLMs) to effectively fuse multimodal features within a unified attention-driven framework.

2.2 Different Architectures for VQA Over the years, a variety of architectures have been proposed for Visual Question Answering (VQA), each designed to address specific challenges such as visual grounding, contextual understanding, and reasoning. These architectures differ in how they extract, represent, and fuse visual and textual features, as well as how they handle complex queries. Broadly, they can be categorized into the following types:

Attention-Based Models The attention mechanism was first introduced by Bahdanau et al. [20] in the context of neural machine translation to allow models to dynamically focus on relevant parts of the input sequence. This concept was later extended to computer vision tasks such as image captioning, where attention helped models selectively focus on important regions of the image while generating descriptive captions.

Inspired by its success, attention mechanisms were adopted in Visual Question Answering (VQA) to align visual features with the question context. Early VQA models used question-guided attention to identify relevant image regions based on the question and image-guided attention to focus on textual segments relevant to the visual input [21]. This bidirectional alignment evolved into co-attention mechanisms, such as hierarchical co-attention [22], which applied attention at multiple levels (word, phrase, and sentence) to enhance multi-granular reasoning.

Further advancements led to the use of stacked attention networks [23], where attention was applied repeatedly to refine focus, commonly referred to as single-hop (a single pass of attention) and multi-hop attention (multiple iterations to refine reasoning). These architectures allowed for more complex inference by progressively narrowing down relevant regions and words.

The bottom-up and top-down attention [24] approach became popular in later models. Here, object-level features (bottom-up) are extracted using object detectors, and attention is applied top-down, guided by the question to focus on semantically meaningful regions.

More recently, attention mechanisms have been transformed through the use of self-attention, a core component of transformer architectures [19]. Self-attention enables the model to learn dependencies and interactions across all elements of the input, both within and across modalities. This paradigm shift has led to powerful Vision-Language Models (VLMs) that integrate vision and language understanding into a single unified architecture, significantly advancing the performance of modern VQA systems.

External Knowledge-Based Models Visual Question Answering (VQA) is a complex task that requires answering questions based on visual content, guided by the information provided in the question. While visual and textual inputs are the primary sources for inference, they are often not sufficient to handle questions that require reasoning beyond what is directly observable in the image. In many real-world scenarios, external knowledge is essential to improve the accuracy and depth of understanding.

Questions like "Why are people holding umbrellas?" cannot be answered correctly based solely on image pixels and question words. To interpret such questions accurately, the model must incorporate commonsense knowledge—for instance, understanding that umbrellas are used for protection against rain or sunlight. This highlights a limitation of many VQA datasets, which are typically designed for a fixed scope, making models prone to generating random or inaccurate answers when queried outside of their training context.

To address this challenge, VQA systems have started integrating external knowledge sources into their architecture. These systems leverage two types of knowledge:

- **Implicit Knowledge**: This is the information that the model learns indirectly from large-scale pretraining on massive datasets. It includes patterns, facts, and language associations that the model internalizes during training. While implicit knowledge provides generalization, it may not cover specific or uncommon facts needed for specialized questions.
- **Explicit Knowledge**: This refers to structured or semi-structured data retrieved from curated knowledge bases. It allows models to query and incorporate relevant factual information that is not present in the image or question. This approach helps in answering questions that require reasoning based on external context.

Several external knowledge sources have been used in VQA research to enhance question answering capabilities, including:

- **ConceptNet** [25]: A semantic network of commonsense knowledge that provides relationships between words and concepts.
- **DBpedia** [26]: A structured version of Wikipedia content, offering factual information in RDF format.
- **WebChild** [27]: A knowledge base that connects words and concepts using web-mined relationships, particularly useful for commonsense and comparative knowledge.
- **YAGO** [28]: A large ontology combining information from Wikipedia, WordNet, and GeoNames to provide hierarchical and relational knowledge.
- **Freebase** [29]: A collaboratively created knowledge graph containing structured facts about people, places, and things, often used for entity-based reasoning.

By integrating these knowledge sources, external knowledge-based VQA models can perform more robust reasoning, make informed inferences, and handle open-domain questions that go beyond visual-textual inputs. This approach bridges

the gap between visual understanding and real-world knowledge, significantly improving performance in complex and explanatory VQA tasks.

Language Bias Mitigation Based Models There is no doubt that Visual Question Answering (VQA) has achieved remarkable results. However, several studies have identified significant language and dataset biases within VQA benchmarks. One of the main reasons is the strong priors present in the training and test splits of the VQA datasets. As a result, when faced with complex reasoning tasks, VQA models often rely on these priors—answering questions based on language correlations rather than actual visual understanding [30].

For example, in the VQA 1.0 dataset, many counting-related questions are disproportionately answered with "2", and most color-related questions tend to have "white" as the answer. To address this issue, the authors of VQA 1.0 revised the dataset by modifying the answer distributions to reduce such biases. Various approaches have been proposed to overcome language biases in VQA systems. One such method involves adversarial regularization [31], which utilizes a dual-stream architecture. The first stream processes both visual and textual data, while the second stream focuses solely on the question, allowing the model to disentangle language priors from visual reasoning.

Another approach is ensemble-based learning [32], which enhances performance on out-of-domain VQA questions by fusing the gradients from multiple networks to improve generalization.

In addition, counterfactual data augmentation has been employed to generate counterfactual samples, helping to reduce bias during training by exposing the model to more diverse question-answer pairs [33].

Furthermore, the integration of Generative Adversarial Network (GAN) objectives, knowledge distillation, and vision-language models (VLMs) has also gained traction, offering improved robustness and fairness in VQA tasks. Overall, these techniques aim to make VQA models less reliant on superficial correlations and more grounded in visual reasoning, leading to fairer and more accurate performance across diverse scenarios [34, 35].

Reasoning Models VQA that requires reasoning over both visual and textual information is often referred to as compositional VQA. These models perform multi-step reasoning on images using cues from the associated questions. For example, consider the question: "What is the color of the object above the oval shape?"—this requires the model to first locate the oval shape, then identify the object positioned above it, and finally classify its color.

To support such complex reasoning, various compositional architectures have been proposed. These include Neural Module Networks (NMNs) [36] and Episodic Memory Modules, which allow dynamic reasoning based on question structure. Another example is the Multi-modal Relational Network (MUReL) [37], which learns representations by reasoning over visual inputs in an end-to-end manner.

A more recent approach is Programmatic VQA (PVQA) [38], which answers questions by generating executable programs based on the question and visual

content. PVQA typically consists of three components: (1) Large Language Models (LLMs) for code generation, (2) Image processing modules to extract relevant features, and (3) a code executor to run the generated programs. The LLM takes the question as input and produces code that, when executed with the help of the defined image processing modules, yields the answer.

These compositional approaches enable fine-grained, interpretable reasoning and are especially effective for complex queries that require structured understanding of visual scenes.

Vision-Language Models (VLMs) In the current era, Large Language Models (LLMs) have demonstrated remarkable capabilities in capturing and applying world knowledge, resulting in exceptional performance across various Natural Language Processing (NLP) tasks [39]. Motivated by these advancements, researchers have extended their investigation to vision-language tasks, leading to the emergence of Vision-Language Models (VLMs). The primary objective of Vision-Language Models (VLMs) is to effectively learn the correlation between image-text pairs. Given a vision-text data pair as input, VLMs extract both visual and textual features and apply a pretraining objective function to model their semantic alignment. Once trained, VLMs are evaluated using a zero-shot learning approach on unseen data by aligning the embeddings of image-text pairs in a shared representation space [40].

VLM pretraining involves the use of deep neural networks to extract visual and textual features from paired image-text data. The architecture of a typical VLM consists of two main components: an image encoder and a text encoder, both based on deep neural networks. These encoders are responsible for transforming the input modalities into corresponding image and text embeddings within a shared representation space.

For visual feature extraction, two main architectures have been widely adopted: Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). Early VLMs primarily relied on CNN-based models such as VGG [10], ResNet [11], and GoogLeNet [12] due to their strong performance in traditional computer vision tasks. However, in recent years, Vision Transformers [13] have gained significant popularity for tasks like image classification, object detection, and segmentation. Most state-of-the-art VLMs now employ Vision Transformers for visual representation learning. On the textual side, Transformer-based architectures [19] and their variants have become the standard choice, owing to their success in modeling complex linguistic patterns and context.

Pretraining objectives play a crucial role in the development of Vision-Language Models (VLMs), as they enable the models to effectively understand and associate visual and textual information. These objectives generally fall into three categories: contrastive, generative, and alignment-based approaches.

Together, these pretraining strategies enable VLMs to capture rich multimodal representations, forming the foundation for their success across a wide range of downstream vision-language tasks like VQA [41].

2.3 Applications of VQA Visual Question Answering (VQA) is increasingly being used in real-world scenarios by combining computer vision and natural language processing [42]. Here are some of its practical applications:

Healthcare and Medical Imaging VQA can assist doctors in analyzing medical images like X-rays or MRIs by answering specific questions, such as detecting fractures or abnormalities. It enhances diagnostic efficiency and supports medical decision-making [6].

Assisting Visually Impaired Individuals VQA enables blind or visually impaired users to interact with their environment by asking questions about images or surroundings, such as identifying objects or reading signs [4].

Education and Advertising In education, VQA can make learning more interactive by helping students understand images. In advertising, it helps analyze and describe visual content, ensuring brand consistency and message clarity [5, 43].

Surveillance and Autonomous Vehicles VQA improves surveillance by helping systems identify unusual activities or objects. In autonomous vehicles, it helps interpret road scenes, such as detecting pedestrians or traffic signals [44].

With advancements in artificial intelligence, the role of Visual Question Answering (VQA) in daily life continues to grow rapidly. In the near future, VQA will be integrated into intelligent chatbots and virtual assistants, enabling users to interact with images and videos more naturally. These systems will not only answer visual questions but also offer real-time guidance, support decision-making, and personalize user experiences across healthcare, education, shopping, and smart home environments. As VQA evolves, it will make technology more accessible, intelligent, and seamlessly responsive to both visual and verbal human queries.

3 Lidar Meets VQA

Light Detection and Ranging (LiDAR) is a remote sensing technology that uses laser pulses to measure distances to objects and generate high-resolution, three-dimensional information about the environment. LiDAR sensors produce point cloud data, which is a dense collection of data points in three-dimensional space. Each point represents a precise location on the surface of an object, typically defined by its X, Y, and Z coordinates. These points are generated when laser pulses emitted by the sensor reflect off surfaces and return to the receiver, allowing accurate measurement of distance and shape. Point clouds capture the geometric structure of real-world environments and are crucial for tasks such as object detection, 3D reconstruction, and scene understanding. The richness of spatial

details in point clouds makes them highly valuable in computer vision applications, especially in autonomous navigation, robotics, and smart infrastructure mapping. Unlike traditional 2D images, LiDAR provides structural geometry, making it a valuable modality for visual understanding in autonomous systems, robotics, and environmental monitoring.

In recent years, significant advancements have been made in vision-language tasks such as Visual Question Answering (VQA) and image captioning, particularly on 2D image data. However, when these models—originally designed for 2D images—are applied to 3D scenes, their performance often degrades. This decline is primarily due to the lack of spatial context, such as relative directions and distances, which are not well represented in 2D projections. Furthermore, occlusion becomes a major challenge, as some objects may be partially or fully hidden behind others, making it difficult for models to accurately identify and localize them. As a result, these models struggle with reliable object localization and identification in complex 3D environments [45, 46].

Currently, several 3D spatial understanding models have been developed for tasks such as 3D object localization. Notable examples include ScanRefer [47], ReferIt3D [48], and Scan2Cap [49], which leverage 3D data to ground language in spatial scenes. However, compared to their 2D counterparts, datasets for 3D Visual Question Answering (VQA) are still limited in both scale and diversity. They often lack rich environmental annotations and offer a narrow range of question types, which restricts the development and evaluation of more generalized and robust 3D VQA models.

3.1 Deep Learning Models for Point Cloud Understanding Models designed to process point cloud data are specifically tailored to handle its unordered, irregular, and sparse structure. Point-based models such as PointNet and PointNet++ directly operate on raw 3D points. PointNet [45] treats each point independently and then aggregates global features, while PointNet++ [50] improves upon this by capturing local neighborhood structures through hierarchical learning. Another notable model, DGCNN (Dynamic Graph CNN) [51], constructs dynamic graphs to capture local geometric relationships using edge convolutions.

Voxel-based models convert point clouds into regular 3D voxel grids and apply 3D convolutional networks. VoxNet [52] was one of the earliest models in this category, followed by more complex architectures like 3D ResNets and 3D U-Nets. Although effective, voxel-based methods often suffer from high memory consumption due to the sparsity of 3D space. To address this, octree-based models such as O-CNN [54] use hierarchical spatial partitioning (octrees) to reduce computational costs while maintaining spatial detail.

Another approach involves projection-based models, which project point clouds onto 2D views and process them with traditional 2D CNNs. Multi-view CNN (MVCNN) is a prominent example that combines multiple 2D projections to represent the 3D structure [59]. More recently, transformer-based and hybrid models have gained attention. For example, the Point Transformer [53] lever-

ages self-attention mechanisms adapted for point clouds to model long-range dependencies, while PAConv introduces dynamic kernel aggregation to improve feature flexibility and learning capacity.

Finally, graph-based models treat point clouds as graphs and apply graph convolutions to capture spatial relationships between neighbouring points [55]. These diverse approaches reflect the growing interest in effectively learning from point cloud data across a variety of complex 3D understanding tasks. However, existing 3D data understanding approaches often struggle to generalize to unseen environments or data distributions. As a result, their performance on downstream tasks such as Visual Question Answering (VQA) and scene captioning remains suboptimal. To address this limitation, recent research has focused on incorporating 3D point cloud data, leveraging its inherent sparsity and rich geometric relationships. By capturing spatial structures more effectively, point clouds offer a more robust foundation for building models capable of understanding complex 3D scenes.

3.2 VQA Using Point Cloud Data Visual Question Answering (VQA) using point cloud data extends traditional 2D VQA into the 3D domain, enabling machines to understand and reason about complex spatial environments. Unlike images, point clouds provide rich geometric and depth information, making them ideal for tasks that require spatial awareness and object localization. This approach empowers applications in robotics, autonomous systems, and assistive technologies that operate in real-world 3D spaces.

Point Cloud QA focuses specifically on leveraging raw 3D point cloud data for spatial reasoning and object-centric queries. Both aim to enrich scene understanding by incorporating depth and geometric context beyond 2D vision.

The proposed work TransVQA3D [56] leverages 3D point clouds combined with instance indicators for question answering. It utilizes a cross-modal Transformer to fuse object features with language inputs. Subsequently, scene graph initialization is performed, and additional edges are integrated from the scene graph to apply scene graph-aware attention for more effective reasoning in PointQA.

Another model, ScanQA [57], addresses question answering in 3D indoor environments using point clouds and sequential RGB frames to capture scene dynamics. The model employs VoteNet to extract geometric features from point clouds and a BiLSTM to encode contextual word features from questions. The final architecture fuses these modalities using three main components: a 3D scene and question encoder, a fusion layer, and a classification layer for answer prediction.

While early research primarily focused on indoor scenes, recent studies have extended these efforts to outdoor urban environments, especially for autonomous driving applications.

The proposed Sg-CityU [58] model tackles this by jointly processing questions, scene graphs, and point clouds. It uses VoteNet as the backbone for point clouds, GCN for scene graphs, and BERT for natural language understanding.

These multimodal features are fused using a combination of self-attention and cross-attention mechanisms in a fusion layer, followed by an answer prediction layer.

Similarly, NuScenes-QA [59] targets autonomous driving scenarios and aims to answer questions based on street-view contextual information. It processes LiDAR point clouds and multi-view images to obtain bird's-eye view (BEV) features, which are then fused with language features to generate answers.

Another approach, AutoQA [44], also focuses on autonomous driving environments, utilizing PointNet++ for extracting features from point clouds and LSTM for processing questions. The fused features are then used for answer generation.

Recently, large language models (LLMs) and vision-language models (VLMs) have demonstrated remarkable performance not only on natural language processing (NLP) tasks but also on multimodal tasks involving 2D image understanding. Building on this progress, researchers have begun to explore their potential for 3D data interpretation. To leverage the capabilities of LLMs in the 3D domain, the proposed work LiDAR-LLM [60] utilizes raw LiDAR data and harnesses the reasoning and understanding power of LLMs to comprehensively interpret complex outdoor 3D scenes.

Overall, these models highlight the growing capability of combining geometric understanding from 3D point clouds with natural language processing to perform complex reasoning tasks in both indoor and outdoor 3D environments.

3.3 Dataset

3.4 ScanQA ScanQA [57] is a 3D question answering (3D-QA) dataset built upon RGB-D scans of indoor scenes, with annotations derived from the ScanNet dataset. The captions were initially generated using object-level captions from ScanRefer and were manually refined to correct any invalid entries. Additionally, the authors collected free-form answers and object annotations from human annotators using an interactive 3D scene viewer. Overall, the dataset comprises 41,362 questions and 58,191 answers, including 32,337 unique questions and 16,999 unique answers.

3.5 CLEVER3D CLEVR3D [56] is a Visual Question Answering (VQA) dataset based on 3D point clouds, designed for VQA-3D tasks. Its foundation lies in the 3D Semantic Scene Graph (3DSSG) annotations [10] from the 3RScan dataset [44]. In total, CLEVR3D comprises 8,771 scenes and 171,174 associated questions. The dataset is divided into two subsets: (a) CLEVR3D-REAL and (b) CLEVR3D-SIM, used to evaluate real-scene VQA-3D and common-sense independent VQA-3D (CSI-VQA-3D), respectively.

For CLEVR3D-REAL, the training and test sets consist of 38,806 and 5,765 questions from 1,176 and 157 real 3D scenes, respectively. In contrast, CLEVR3D-SIM includes 109,351 training questions and 17,252 test questions derived from 6,510 and 928 simulated 3D scenes.

3.6 nuScenes-QA NuScenes-QA [59] presents a dataset for Visual Question Answering (VQA) in the context of autonomous driving, designed to answer queries based on street-view clues. The dataset is built upon the nuScenes dataset, a widely used benchmark for autonomous driving research. Inspired by the CLEVR dataset, NuScenes-QA adopts a similar approach to annotate question–answer pairs.

Overall, NuScenes-QA comprises 459,941 question–answer pairs related to 34,149 visual scenes. The training split includes 376,604 questions from 28,130 scenes, while the testing split contains 83,337 questions from 6,019 scenes.

3.7 City-3DQA City-3DQA [58] supports city-level scene understanding for VQA by incorporating scene semantics and human–environment interaction tasks within urban environments. It is built upon the 3D city point cloud dataset UrbanBIS. City-3DQA covers 193 unique scenes across six cities: Qingdao, Wuhu, Longhua, Yuehai, Lihu, and Yingrenshi, incorporating a total of 2.5 billion point clouds.

The dataset includes information from 3,370 instances across these cities, covering elements such as buildings, vehicles, bridges, and boats.

3.8 Auto-QA The AutoQA [44] dataset aims to investigate and evaluate the effectiveness of localization-based question answering, specifically in the context of autonomous driving—where such functionality is particularly critical. This dataset is built on top of the Argoverse dataset and offers a multimodal setting, providing seven views per frame along with point cloud data to support answering localization-related questions.

Overall, the dataset contains 31,259 scenes associated with 300,000 questions. These questions cover 15 different object categories present in the scenes and include 4,000 unique question types.

3.9 Future Recommendations The fusion of LiDAR point clouds with language models for VQA is still an emerging area. With rapid progress in large language and vision models, several research directions are recommended for future exploration:

Foundation Model Integration for VQA-3D The emergence of vision-language models (VLMs) such as GPT-4V [61] and LLaVA [62] has opened up new possibilities for multimodal learning. Future research can harness these models to enable zero-shot and few-shot VQA directly on 3D LiDAR scenes without heavy supervision.

Real-Time VQA for Robotics and Autonomous Vehicles Future models should be optimized for low-latency processing to support real-time VQA in robotics, drones, and self-driving systems using sparse LiDAR data and minimal compute.

Interactive and Multi-turn VQA in 3D Enabling dialogue-based 3D VQA for real-world tasks—like navigation, robot assistance, or situational analysis—can make systems more human-centric and useful in dynamic environments.

Enhanced 3D Scene Understanding through Scene Graphs Future systems should incorporate semantic scene graphs for complex spatial reasoning, occlusion handling, and relational understanding of 3D environments in response to natural language queries.

4 Conclusion

This chapter provided a comprehensive overview of the evolution and integration of Visual Question Answering (VQA) with LiDAR data for enhanced multimodal scene understanding. We began by exploring the fundamentals of VQA, its core architecture, and the major categories of models, including attention-based methods, external knowledge integration, language bias mitigation strategies, reasoning-based approaches, and emerging vision-language models (VLMs).

Following this, we delved into the characteristics of LiDAR sensors, focusing on their ability to generate rich 3D point cloud data. We discussed how deep learning techniques have been adapted to process point clouds, and examined their role in extending VQA to 3D spatial understanding. Several benchmark datasets supporting point cloud-based VQA were reviewed, highlighting their contributions and limitations.

Finally, we emphasized the future potential of combining LiDAR data with powerful language and vision models. Recommendations were made for advancing this field, including leveraging foundation models, enhancing spatial reasoning, and enabling real-time, context-aware applications in robotics and autonomous driving. As multimodal AI continues to evolve, the fusion of LiDAR and VQA offers promising pathways for deeper, geometry-aware human-machine interaction.

Bibliography

[1] Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision 2015 (pp. 2425-2433).

[2] Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 2625-2634).

[3] Noh H, Araujo A, Sim J, Weyand T, Han B. Large-scale image retrieval with attentive deep local features. In Proceedings of the IEEE international conference on computer vision 2017 (pp. 3456-3465).

[4] Gurari D, Li Q, Stangl AJ, Guo A, Lin C, Grauman K, Luo J, Bigham JP. Vizwiz grand challenge: Answering visual questions from blind people. In Proceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 3608-3617).

[5] Das A, Kottur S, Gupta K, Singh A, Yadav D, Moura JM, Parikh D, Batra D. Visual dialog. In Proceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 326-335).

[6] Ben Abacha A, Hasan SA, Datla VV, Demner-Fushman D, Müller H. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes 2019. 9-12 September 2019.

[7] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition 2009 Jun 20 (pp. 248-255). Ieee.

[8] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft coco: Common objects in context. In Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13 2014 (pp. 740-755). Springer International Publishing.

[9] Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results [Internet]. 2012 [cited 2025 Jun 6]. Available from: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html

[10] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014 Sep 4.

[11] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-778).

[12] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 1-9).

[13] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020 Oct 22.
[14] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems. 2015;28.
[15] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013 Jan 16.
[16] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997 Nov 15;9(8):1735-80.
[17] Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078. 2014 Jun 3.
[18] Devlin J, Chang MW, 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) 2019 Jun (pp. 4171-4186).
[19] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017;30.
[20] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473. 2014 Sep 1.
[21] Zhu Y, Groth O, Bernstein M, Fei-Fei L. Visual7w: Grounded question answering in images. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 4995-5004).
[22] Lu J, Yang J, Batra D, Parikh D. Hierarchical question-image co-attention for visual question answering. Advances in neural information processing systems. 2016;29.
[23] Yang Z, He X, Gao J, Deng L, Smola A. Stacked attention networks for image question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 21-29).
[24] Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 6077-6086).
[25] Speer R, Chin J, Havasi C. Conceptnet 5.5: An open multilingual graph of general knowledge. In Proceedings of the AAAI conference on artificial intelligence 2017 Feb 12 (Vol. 31, No. 1).
[26] Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z. Dbpedia: A nucleus for a web of open data. In international semantic web conference 2007 Nov 11 (pp. 722-735). Berlin, Heidelberg: Springer Berlin Heidelberg.
[27] Tandon N, De Melo G, Suchanek F, Weikum G. Webchild: Harvesting and organizing commonsense knowledge from the web. In Proceedings of the 7th

ACM international conference on Web search and data mining 2014 Feb 24 (pp. 523-532).
[28] Suchanek FM, Kasneci G, Weikum G. Yago: a core of semantic knowledge. In Proceedings of the 16th international conference on World Wide Web 2007 May 8 (pp. 697-706).
[29] Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J. Freebase: a collaboratively created graph database for structuring human knowledge. InProceedings of the 2008 ACM SIGMOD international conference on Management of data 2008 Jun 9 (pp. 1247-1250).
[30] Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 6904-6913).
[31] Ramakrishnan S, Agrawal A, Lee S. Overcoming language priors in visual question answering with adversarial regularization. Advances in neural information processing systems. 2018;31.
[32] Clark C, Yatskar M, Zettlemoyer L. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. arXiv preprint arXiv:1909.03683. 2019 Sep 9.
[33] Niu Y, Tang K, Zhang H, Lu Z, Hua XS, Wen JR. Counterfactual vqa: A cause-effect look at language bias. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2021 (pp. 12700-12710).
[34] Cho JW, Kim DJ, Ryu H, Kweon IS. Generative bias for robust visual question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2023 (pp. 11681-11690).
[35] Rahman T, Chou SH, Sigal L, Carenini G. An improved attention for visual question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2021 (pp. 1653-1662).
[36] Andreas J, Rohrbach M, Darrell T, Klein D. Neural module networks. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 39-48).
[37] Cadene R, Ben-Younes H, Cord M, Thome N. Murel: Multimodal relational reasoning for visual question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2019 (pp. 1989-1998).
[38] Gupta T, Kembhavi A. Visual programming: Compositional visual reasoning without training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023 (pp. 14953-14962).
[39] Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le QV, Zhou D. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems. 2022 Dec 6;35:24824-37.
[40] Lu J, Batra D, Parikh D, Lee S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems. 2019;32.
[41] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G. Learning transferable visual models from natural language supervision. In International conference on machine learning 2021 Jul 1 (pp. 8748-8763). PmLR.

[42] Barra S, Bisogni C, De Marsico M, Ricciardi S. Visual question answering: Which investigated applications?. Pattern Recognition Letters. 2021 Nov 1;151:325-31.

[43] Stefanini M, Cornia M, Baraldi L, Corsini M, Cucchiara R. Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain. InImage Analysis and Processing–ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part II 20 2019 (pp. 729-740). Springer International Publishing.

[44] Kumar S, Patro BN, Namboodiri VP. Auto qa: The question is not only what, but also where. InProceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 2022 (pp. 272-281).

[45] Chen X, Ma H, Wan J, Li B, Xia T. Multi-view 3d object detection network for autonomous driving. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition 2017 (pp. 1907-1915).

[46] Qi CR, Su H, Mo K, Guibas LJ. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 652-660).

[47] Chen DZ, Chang AX, Nießner M. Scanrefer: 3d object localization in rgb-d scans using natural language. In European conference on computer vision 2020 Aug 23 (pp. 202-221). Cham: Springer International Publishing.

[48] Achlioptas P, Abdelreheem A, Xia F, Elhoseiny M, Guibas L. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16 2020 (pp. 422-440). Springer International Publishing.

[49] Chen Z, Gholami A, Nießner M, Chang AX. Scan2cap: Context-aware dense captioning in rgb-d scans. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2021 (pp. 3193-3203).

[50] Qi CR, Yi L, Su H, Guibas LJ. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems. 2017;30.

[51] Phan AV, Le Nguyen M, Nguyen YL, Bui LT. Dgcnn: A convolutional neural network over large-scale labeled graphs. Neural Networks. 2018 Dec 1;108:533-43.

[52] Maturana D, Scherer S. Voxnet: A 3d convolutional neural network for real-time object recognition. In 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS) 2015 Sep 28 (pp. 922-928). IEEE.

[53] Zhao H, Jiang L, Jia J, Torr PH, Koltun V. Point transformer. In Proceedings of the IEEE/CVF international conference on computer vision 2021 (pp. 16259-16268).

[54] Wang PS, Liu Y, Guo YX, Sun CY, Tong X. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. ACM Transactions On Graphics (TOG). 2017 Jul 20;36(4):1-1.

[55] Zhou H, Feng Y, Fang M, Wei M, Qin J, Lu T. Adaptive graph convolution for point cloud analysis. In Proceedings of the IEEE/CVF international conference on computer vision 2021 (pp. 4965-4974).

[56] Yan X, Yuan Z, Du Y, Liao Y, Guo Y, Cui S, Li Z. Comprehensive visual question answering on point clouds through compositional scene manipulation. IEEE Transactions on Visualization and Computer Graphics. 2023 Dec 8;30(12):7473-85.
[57] Azuma D, Miyanishi T, Kurita S, Kawanabe M. Scanqa: 3d question answering for spatial scene understanding. Inproceedings of the IEEE/CVF conference on computer vision and pattern recognition 2022 (pp. 19129-19139).
[58] Sun P, Song Y, Liu X, Yang X, Wang Q, Li T, Yang Y, Chu X. 3d question answering for city scene understanding. In Proceedings of the 32nd ACM International Conference on Multimedia 2024 Oct 28 (pp. 2156-2165).
[59] Qian T, Chen J, Zhuo L, Jiao Y, Jiang YG. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In Proceedings of the AAAI Conference on Artificial Intelligence 2024 Mar 24 (Vol. 38, No. 5, pp. 4542-4550).
[60] Yang S, Liu J, Zhang R, Pan M, Guo Z, Li X, Chen Z, Gao P, Li H, Guo Y, Zhang S. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. In Proceedings of the AAAI Conference on Artificial Intelligence 2025 Apr 11 (Vol. 39, No. 9, pp. 9247-9255).
[61] OpenAI. GPT-4V Technical Report. 2023. Available from: https://openai.com/research/gpt-4v
[62] Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. Advances in neural information processing systems. 2023 Dec 15;36:34892-916.