International Journal of Informatics and Communication Technology (IJ-ICT)
Vol. 15, No. 3, September 2026, pp. 1097~1114
ISSN: 2252-8776, DOI: 10.11591/ijict.v15i3.pp1097-1114

Exploring multi-answer visual question answering with object detection: a systematic review

Nidaul Hasanati[1,2], Taufik Djatna[3], Imas Sukaesih Sitanggang[4], Arif Imam Suroso[5]
[1]Doctoral Program in Computer Science, School of Data Science, Mathematics, and Informatics, IPB University, Bogor, Indonesia
[2]Department of Information Systems, Syarif Hidayatullah State Islamic University (UIN), Jakarta, Indonesia
[3]Division of Agroindustrial Engineering and Technology, Faculty of Engineering and Technology, IPB University, Bogor, Indonesia
[4]School of Data Science, Mathematics, and Informatics, IPB University, Bogor, Indonesia
[5]School of Business, IPB University, Bogor, Indonesia

Article Info

Article history:

Received Jan 26, 2026
Revised Jun 2, 2026
Accepted Jul 5, 2026

Keywords:

Multi-answer
Multi-label
Multiple answers
Object detection
Systematic review
Visual question answering

ABSTRACT

Visual question answering (VQA) is a challenging research area that enables machines to answer natural language questions based on visual content by jointly understanding images and text. Conventional VQA systems typically produce a single answer for each image–question pair. However, many real-world visual questions are ambiguous or complex, allowing multiple valid answers to exist. This systematic literature review (SLR) focuses on multi-answer VQA systems and the use of object detection, following the PRISMA 2020 guidelines. We analyzed 58 peer-reviewed journal articles retrieved from the Scopus database published between 2020 and 2025. Ten of these studies clearly stated that generating multiple answers was their main goal. Forty-eight others indirectly supported answer variability by using object-based or multi-instance reasoning. Through this review, we examine the current methodologies for supporting multi-answer generation, including model architecture, datasets, and evaluation metrics. Most multi-answer generation approaches utilize attention mechanisms, graph neural networks, and transformer-based models. Additionally, we propose a taxonomy of multi-answer VQA organized along four dimensions. Limitations are identified in datasets and evaluation metrics (i.e., answer ambiguity/subjectivity). Future research should focus on improving model interpretability and designing an evaluation framework that incorporates subjective and context-sensitive responses.

Corresponding Author:

Taufik Djatna
Division of Agroindustrial Engineering and Technology, Faculty of Engineering and Technology
IPB University
Bogor, Indonesia
Email: taufikdjatna@apps.ipb.ac.id

1. INTRODUCTION

Visual question answering (VQA) represents a multidisciplinary research domain that converges computer vision and natural language processing to enable machines to comprehend and respond to natural language queries based on visual content [1]. This rapidly evolving field has witnessed substantial advancement through the integration of deep learning technologies, the growth of large-scale multimodal datasets, and the development of sophisticated vision-language approaches across diverse application domains, including healthcare, robotics, remote sensing, and artificial intelligence systems [2], [3].

Despite these significant developments, conventional VQA systems predominantly operate under the restrictive assumption of single-answer generation, wherein each visual question is mapped to precisely

one correct response. The single-answer paradigm proves fundamentally inadequate when faced with ambiguous visual realities, where a single image often holds multiple layers of truth that cannot be summarized in a single label [4]. Recent research shows that VQA systems should be able to provide multiple correct answers for multiple objects in a complex visual scene, ambiguous questions, and/or questions having subjective answers [5].

The emergence of multi-answer VQA systems addresses several fundamental limitations in traditional approaches. Today's visual content is composed of complex scenes with multiple objects, varying attributes, and complex relationships between them, all of which may create situations where multiple correct answers may validly apply to a single question. For example, if an image contains multiple shades of banana color, several responses could indicate what color banana is present in the image. Additionally, questions regarding human emotions or human activity could have different interpretations of what is occurring based on a user's culture, their social context, and their individual experiences [6]. As VQA systems become increasingly interactive, comprehensive answers covering multiple valid interpretations of visual content become essential.

Object detection is an essential technology for the VQA system to generate multiple answers because it helps create a base visual understanding of the scene and identify, locate, and define multiple entities or objects in a complex environment. When used with object-based representation models, VQA systems can systematically analyze the visual content, correlate the identified objects, and produce multiple answers for a given question based on rich visual data [7]. Without precise object detection, a VQA system operates without explicit visual grounding. Object detection integration serves as a visual anchor that allows the system to granularly dissect each entity, help clarify ambiguous questions, and enable the creation of a complete set of answers that fully represent the complexity of visual content.

From an information and communication technology (ICT) perspective, multi-answer VQA systems introduce several computational and communication challenges. Modern VQA pipelines typically rely on cloud-based processing of high-resolution images combined with multimodal deep learning models, which require substantial computational power and data transmission. In real-world deployment scenarios, edge devices must transmit visual data to a central server for inference, resulting in significant latency and bandwidth challenges for time-critical applications such as medical diagnostic systems, smart agriculture monitoring, and intelligent surveillance systems. Therefore, developing VQA architectures that optimize the balance between inference accuracy and computational or communication overhead is critical for practical ICT deployments [8], [9].

Historically, VQA research has progressed rapidly due to advances in deep learning and multimodal learning. Early studies primarily focused on single-answer prediction, where each image–question pair was assumed to have one correct response. The foundational VQA dataset was introduced alongside CNN-LSTM models as the baseline framework for jointly processing visual and textual information [1]. However, researchers soon recognized that expecting one correct answer to a visual question was often unrealistic, as human annotators could provide different yet equally valid responses to the same question. It has been demonstrated that disagreement among human annotators is frequent in VQA annotations, confirming that multiple valid answers may exist and that models must be able to handle this variability [6]. Furthermore, three primary factors contributing to answer multiplicity have been identified: visual complexity, semantic ambiguity, and subjective interpretation [4].

Subsequent architectural innovations have substantially improved visual reasoning in VQA systems. One of the first explicit frameworks for multi-answer VQA was proposed by introducing mechanisms to generate more than one answer for a single visual question [8]. In parallel, object detection has emerged as an essential technology for VQA. It has been demonstrated that applying Faster R-CNN for object-based visual representation significantly improves visual grounding and reasoning performance [10]. Recent large vision–language models, including BLIP-2 [11] and Flamingo [12], have further advanced multimodal reasoning capabilities by integrating large language models with rich visual representations. Domain-specific VQA applications (including medical imaging, agriculture, and remote sensing) have further emphasized the necessity of multi-answer reasoning, as questions in these domains often require multiple attributes or observations to be described simultaneously [13], [14].

Despite significant progress across these directions, the literature still lacks a systematic synthesis specifically focused on multi-answer VQA systems or the role of object-level visual representations in supporting answer multiplicity. Most existing surveys analyze VQA architectures from the perspective of model evolution or multimodal fusion strategies, including transformer-based models [15], graph reasoning frameworks [16], [17], and domain-specific implementations [18]–[20], without examining how current systems handle situations in which multiple valid answers may exist. This gap motivates the present systematic review. To the best of our knowledge, this review is one of the first to systematically analyze the relationship between object-level visual grounding and answer multiplicity in multi-answer VQA systems.

The main research questions guiding this systematic review are as follows. RQ1: What existing methodological approaches enable multiple answer generation in contemporary VQA systems? RQ2: How does object detection contribute to reasoning capability and answer multiplicity in VQA models? RQ3: What are the datasets and evaluation metrics commonly used in studies that combine object detection and multi-answer VQA? RQ4: What are the principal limitations, persistent challenges, and promising future research directions identified in the literature regarding object-based multi-answer VQA systems?

The concept of multi-answer VQA remains inconsistently defined across studies. Table 1 summarizes common formulations of single-answer and multi-answer VQA tasks, which differ in how the model represents the answer space and handles multiple valid answers. Let an image be denoted as I, a question as Q, an answer as a, and an answer space as $\mathcal{A}$.

Table 1. Formal formulations of VQA tasks related to multi-answer reasoning

Formulation	Mathematical expression	Description
Single-answer VQA	$f:(I,Q) \rightarrow a$, where $a \in \mathcal{A}$	Each image–question pair is mapped to one dominant answer (SoftMax output layers).
Multi-label VQA	$f:(I,Q) \rightarrow \{a_1, a_2, \ldots, a_k\}$	Multiple answers predicted simultaneously using multi-label classification (sigmoid output layers).
Generative multi-answer	$f:(I,Q) \rightarrow \text{sequence}(a_1, a_2, \ldots, a_k)$	The model generates a sequence of plausible answers based on conditional probability distributions.
Top-k candidate ranking	$f:(I,Q) \rightarrow \text{Top-}k(\mathcal{A})$	The model ranks candidate answers and returns several top predictions from a predefined answer space (multiple-choice VQA).
Implicit multi-answer reasoning	$f:(I,Q,O_1,\ldots,O_n) \rightarrow a$	Model reasons over multiple detected objects but outputs a single answer (object-based VQA).

Based on these formulations, this review regards a study as supporting multi-answer VQA when at least one of the following conditions is satisfied: (i) the model explicitly predicts multiple answers, such as through multi-label or answer-set outputs; (i) the model generates several semantically distinct responses for the same image–question pair; or (iii) the model employs object-level or region-based reasoning mechanisms that enable interpretation of multiple visual entities.

The main goals of this systematic review are to: (i) provide an organized summary of recent advances in multi-answer VQA systems incorporating object detection; (ii) categorize and compare state-of-the-art deep learning methods based on architectural design and multimodal fusion strategies; (iii) overview commonly used datasets and evaluation metrics relevant to multi-answer or object-aware VQA systems; and (iv) identify key challenges and future research directions for more robust and interpretable multi-answer VQA models. To support these goals, this review further contributes a structured four-dimensional taxonomy of multi-answer VQA, organizing existing approaches along task formulation, visual representation strategy, sources of answer multiplicity, and evaluation strategy dimensions, with particular emphasis on how object-level visual grounding enables answer multiplicity. The remainder of this paper is organized as follows. Section 2 describes the research methodology. Section 3 presents the results and discussion. Section 4 concludes the paper.

2. RESEARCH METHOD

This systematic review employed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines to ensure a comprehensive, transparent, and reproducible investigation [21]. The completed PRISMA 2020 checklist is provided in Supplementary Material, Table S1.

2.1. Search strategy and information resources

The search strategy combined three keyword groups: the VQA task (“visual question answering”), multi-answer concepts (“multiple answer”, “multi-answer”, “multi-label”), and object-level visual representations (“object detection”, “region”, “bounding box”). These terms were selected to capture both explicit multi-answer approaches and studies that implicitly support answer variability through object-level reasoning. The keyword categories and Boolean search strings used in the query are summarized in Table 2.

The search was conducted in January 2026 using the Scopus database, which was selected for its extensive coverage of peer-reviewed journals in the fields of computer vision and artificial intelligence, including publications from IEEE, ACM, Elsevier, and Springer. Filters were applied to limit results to English-language journal articles published between January 2020 and December 2025.

Technical architectural terms such as “object detection,” “region,” and “bounding box” were prioritized over abstract semantic terms to enhance recall of studies that implicitly address answer variability

through object-level reasoning without explicitly using ambiguity terminology. This methodological decision is because in the VQA domain, question ambiguity is often considered a conceptual phenomenon, while technical solutions are implemented through visualization and entity localization mechanisms.

Table 2. Keyword categories and search terms, and Boolean search strings

Keyword category	Search terms	Rationale
Core task	"Visual question answering"	Used in full form to avoid ambiguity with other meanings of the acronym VQA, such as video quality assurance.
Multi-answer concept	"Multiple answer", "multi-answer", "multi answer", "multi-label"	To identify VQA studies that allow more than one valid answer for a single image–question pair.
Object-level visual representation	"Object detection", "region", "bounding box"	To retrieve studies that support multi-object and multi-answer reasoning through object-level visual grounding, even when the term "multiple answer" is not explicitly used.
Scopus	TITLE-ABS-KEY ("visual question answering") AND (("multiple answer" OR "multi-answer" OR "multi answer" OR "multi-label") OR ("object detection" OR "region" OR "bounding box").	Journals, English, 2020-2025

2.2. Eligibility criteria

Studies were included if they were published in a peer-reviewed journal indexed in Scopus, written in English, published between January 2020 and December 2025, and directly addressed VQA tasks while either explicitly endorsing multi-answer formulations or implicitly facilitating multiple valid answers through object-based, region-based, or multi-instance reasoning mechanisms. The restriction to peer-reviewed journal articles ensures a minimum quality threshold consistent with established SLR practices in applied computer science [22]. Studies were excluded if they lacked experimental validation or reliable evaluation procedures, did not focus on VQA tasks, or did not employ object-based or region-based reasoning to support multiple valid answers. Survey papers, review articles, and other secondary studies were also excluded. Future reviews may consider including high-impact conference proceedings to expand scope.

2.3. Selection process

The selection process followed four sequential PRISMA phases: (i) Identification, in which records were retrieved from Scopus using the predefined search string; (ii) Screening, in which two authors independently screened titles and abstracts against inclusion and exclusion criteria; (iii) Eligibility assessment, in which full-text articles were evaluated independently by the same reviewers, with disagreements resolved through discussion and additional author consultation when needed; and (iv) Final inclusion, in which studies satisfying all criteria were included in the final synthesis. The complete process is summarized in the PRISMA flow diagram as shown in Figure 1. Although a formal inter-rater reliability coefficient (e.g., Cohen's Kappa) was not calculated prior to consensus resolution, the use of clearly defined eligibility criteria and independent multi-stage screening was intended to minimize subjectivity in the selection process, consistent with established SLR practices in artificial intelligence and computer vision research [22].

2.4. Data extraction and quality assessment

Data extraction captured descriptive characteristics (application domain, research objectives, dataset utilization), methodological details (evaluation metrics, object-level visual representation strategies, explicit or implicit multi-answer support), and study-level observations (contributions, limitations, future directions). Quality assessment was adapted for AI-based VQA research using five criteria: (i) clarity of problem formulation, (ii) transparency of dataset descriptions, (iii) appropriateness of object detection integration, (iv) clarity of evaluation protocols, and (v) completeness of reported results.

Each criterion was independently rated as low, moderate, or high risk of bias by two reviewers, with disagreements resolved through consensus. Most studies received low risk ratings across all five criteria, particularly those that used well-known public datasets and reported their evaluation methods clearly and completely. Studies that relied on custom datasets without public access or that provided incomplete descriptions of their evaluation procedures were assigned moderate risk ratings. High risk of bias was not identified in any of the included studies, as all had undergone peer-review validation prior to inclusion. Quality ratings were used to contextualize findings during narrative synthesis rather than as grounds for exclusion.

2.5. Synthesis method

The synthesis followed a narrative and thematic approach in accordance with the PRISMA 2020 guidelines. Narrative synthesis was used to qualitatively summarize and interpret findings from the included studies, while thematic synthesis organized the studies into conceptual categories based on key methodological characteristics relevant to multi-answer VQA research.

Studies were grouped based on methodological characteristics relevant to the research objectives, including task formulation for multi-answer VQA, object-level visual representation strategies, and dataset usage and evaluation protocols. These groupings were defined prior to the synthesis process to ensure consistency across the included studies. Extracted data were subsequently standardized to address variations in terminology and reporting styles.

The characteristics of individual studies were summarized in the study characteristics section. A narrative synthesis was then conducted to identify recurring methodological patterns across the reviewed studies. This approach was selected because substantial heterogeneity in task definitions, datasets, and evaluation metrics precluded the use of quantitative meta-analysis.

3. RESULTS AND DISCUSSION

3.1. Study selection process

The database search identified 600 records from Scopus. After applying filters for publication year, language, and document type, 406 records were removed, leaving 194 for title and abstract screening. Of these, 135 were excluded based on predefined criteria, retaining 59 articles for full-text assessment, as summarized in Table 3. Following full-text evaluation, one additional study was excluded, yielding a final corpus of 58 studies. Among these, 10 explicitly formulated multi-answer VQA as their primary objective, while 48 implicitly supported multi-answer reasoning through object-based, region-based, or multi-instance approaches, as shown in Table 4 and Figure 1.

Table 3. Reasons for exclusion during abstract screening

Reason for exclusion	Number of papers
Duplicate records identified during screening	4
Survey/review articles	4
Insufficient validation or VQA not primary focus	2
Not VQA task	34
VQA has good object detection but a single answer only	62
Single-answer VQA without object-based or multi-instance reasoning	29
Total excluded abstract	135

Table 4. Final study classification

Screening outcome	Number of studies
Explicit multi-answer	10
Implicit multi-answer (object-based, region-based, or multi-instance reasoning)	48
Studies that met all eligibility criteria	58

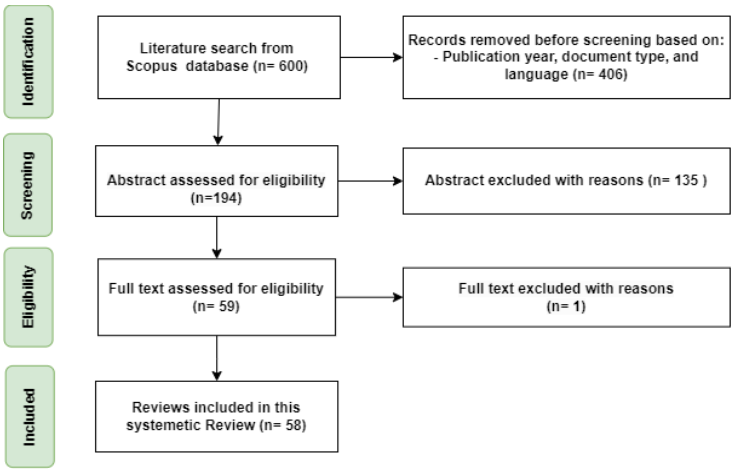


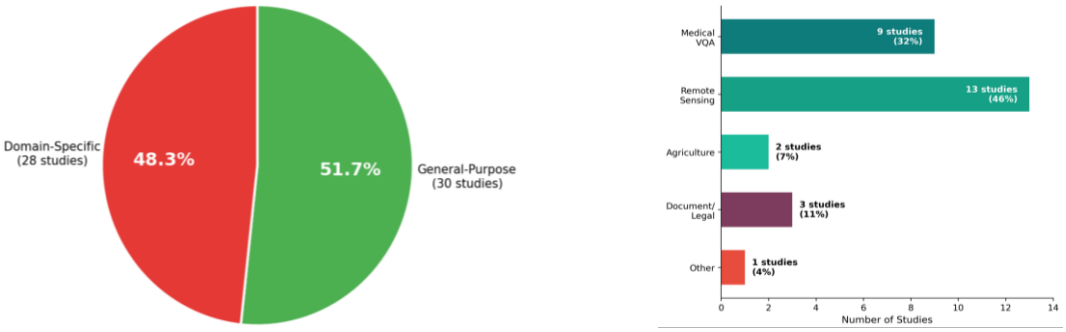
Figure 1. PRISMA flow diagram

3.2. Study characteristics

This section describes the main characteristics of the 58 included studies across four key dimensions: 1) general-purpose versus domain-specific categorization, 2) explicit versus implicit multi-answer formulation, 3) dataset utilization, and 4) object-based visual representation strategies. Detailed characteristics for all 58 included studies are provided in Table S2 (n=28 domain-specific studies) and Table S3 (n=30 general-purpose studies) in Supplementary Material.

3.2.1. General-purpose vs. domain-specific VQA

From the 58 included studies, 30 focus on general-purpose VQA, aiming to answer open-ended questions across diverse everyday images, while the remaining 28 address domain-specific applications, including medical imaging, remote sensing, agriculture, and document analysis. Figure 2 presents the overall distribution of the included studies. Specifically, Figure 2(a) compares the proportions of general-purpose and domain-specific VQA studies, whereas Figure 2(b) illustrates the distribution of domain-specific application areas. General-purpose studies are categorized by task type (e.g., knowledge-based, relational, or open-ended reasoning), while domain-specific studies are categorized by application domain (e.g., medical imaging, remote sensing, or agriculture), as summarized in Table 5.



(a) (b)

Figure 2. Overall study distribution of multi-answer VQA (n=58) (a) general-purpose vs. domain-specific VQA and (b) distribution of domain-specific VQA applications (n = 28)

Table 5. Distribution of included studies

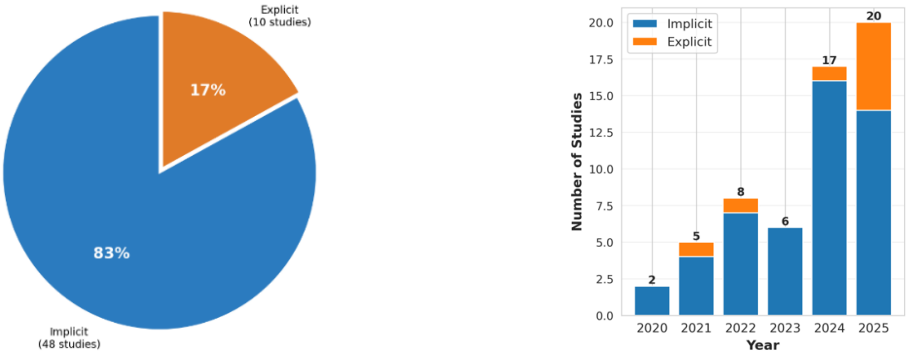
Domain VQA	Num. of studies	Studies
Domain-specific VQA (n=28)		
Medical	9	[23]–[31]
Remote sensing	13	[32]–[44]
Agriculture	2	[14], [45]
Document/Legal	3	[46]–[48]
Other (Facial analysis)	1	[49]
General purposes VQA (n=30)		
General open-ended reasoning	13	[50]–[62]
Relational and spatial reasoning	7	[63]–[69]
Knowledge-based VQA (KB-VQA)	5	[70]–[74]
Specialized tasks (counting and debiasing, explainability)	3	[75]–[77]
Multi-answer and multi-choice design	2	[8], [78]

3.2.2. Explicit vs. implicit multi-answer formulations

Explicit multi-answer studies directly design models to generate multiple answers through multi-label prediction, structured outputs, or generative mechanisms. In contrast, most of the reviewed literature (83%, n=48) supports implicit multi-answer reasoning, in which models incorporate object-based, region-based, or multi-instance mechanisms even though the final prediction remains a single answer. In these models, object detection enables differentiation and comparison of multiple visual entities before generating a response, establishing the architectural foundation for answer variability.

A critical distinction exists between object detection used for multi-entity reasoning versus detection used merely to enhance single-answer accuracy as shown in Figure 3. Among the 135 excluded studies, 62 employed object detection only for single-answer enhancement and were therefore excluded. As shown in Figure 3(a), only 10 studies (17%) explicitly formulated multi-answer VQA, while the remaining 48 (83%)

addressed answer variability implicitly through object-based reasoning mechanisms. Although implicit approaches do not explicitly generate multiple outputs, they are included because object-centric reasoning provides the architectural foundation required for future multi-answer extensions. Figure 3(b) reveals an upward publication trend, with 2024–2025 accounting for 63.8% of included studies (n=37). Notably, 8 of the 10 explicit multi-answer studies were published in this period, reflecting growing awareness of implicit approach limitations. Ten studies explicitly formulated multi-answer generation as their primary objective. These include domain-specific applications in pest management, disaster interpretation, facial expression analysis, and medical VQA, as well as general-purpose approaches addressing multiple-choice VQA direct multiple-answer generation, causality-guided reasoning, and knowledge-augmented VQA as detailed in Table 6.



(a) (b)

Figure 3. Multi-answer VQA formulation distribution (n=58) (a) number of explicit vs implicit studies and (b) publication trends by year and multi-answer type

Table 6. Explicit multi-answer VQA formulations

ID	Study	VQA type	Object detection	Dataset	Key contribution	Output mechanism
1	Hosseinabad *et al.* EMQA [8]	Open-ended	Sliding window	Icon Question Answer (ICQA), DAQUAR	94% less operations, real-time suitable	SoftMaxUnion aggregation
2	Rahnemoonfar *et al.* RescueADI [44]	Remote sensing	bounding boxes	Rescue Adaptive Disaster Interpretation (ADI)	Accuracy 9% higher than existing VQA methods	Structured output tuples (P, R, A)
3	Liu *et al.* SCAG [70]	Knowledge-based	Faster R-CNN	VQAv2, OK-VQA, AK (Ambiguous Knowledge)	Accuracy: VQAv2 = 71.0, OK-VQA = 54.36, AK = 50.42; 4.62% improvement	Multi-target restriction loss
4	Li *et al.* HortiVQA-PP [45]	Horticulture (Pest management)	Semantic segmentation, pest-predator co-occurrence detection	HortiVQA-PP (custom multimodal dataset (30 pest + 10 predator categories))	Multi-label accuracy, Precision, Recall, F1-score, IoU, Hamming Loss, Macro-F1, BLEU-4, ROUGE-L, METEOR, EM, GPT score	Sigmoid activation heads, multi-label matching
5	Zhang *et al.* HiCA-VQA [24]	Medical VQA	Region-based with Hierarchical Modeling + Cross-Attention	Rad-Restruct	F1-score 18% improvement	Hierarchical Answer Decoder (Sigmoid)
6	Upadhyay and Tripathy [73]	Multiple choice VQA	Image regions with attention	Visual7W, VQAv2	Competitive performance	Single-best (Top-4 refinement)
7	Miao *et al.* GAT2R [74]	CGCN (causality-guided)	Object + causal graph	VQA 2.0, VQA v1	improved performance	Single-best selection (SoftMax)
8	Liu *et al.* M2TVQA [63]	Knowledge-augmented VQA	Multi-granularity features/multi-region selection	OK-VQA, KRVQA	Effective verification	Triplet reasoning
9	Liu *et al.* ICQ-TransE [78]	Knowledge-based VQA	Top-K regions (Dynamic visual region selection)	OK-VQA, KRVQA	LLM-enhanced caption bridge	Knowledge graph completion
10	Yuan *et al.* Exp-VQA [49]	Facial expression analysis	InstructBLIP-based visual features, facial regions analysis	Synthesized VQA pairs (emotion classification + Action Unit (AU) detection)	State-of-the-art zero-shot AU detection	Generative narrative captioning

3.2.3. Dataset utilization

A variety of public and domain-specific benchmark datasets are employed across explicit multi-answer VQA studies. As summarized in Table 7, commonly used general-purpose VQA datasets include VQA v2, GQA, OK-VQA, and Visual7W, which are widely adopted for open-domain visual reasoning tasks. In addition, several domain-specific datasets have emerged to support specialized applications, such as RescueADI for remote sensing disaster interpretation, Rad-ReStruct for structured medical report generation, and HortiVQA-PP for agricultural pest management. Furthermore, for fine-grained facial expression analysis, studies like Exp-VQA leverage synthesized VQA pairs derived from established emotion and Action Unit (AU) datasets, including AffectNet, BP4D, DISFA, and RAF-AU, to enable multi-scale reasoning from macro to micro expressions.

Only a limited number of datasets explicitly support multi-answer reasoning. Datasets such as DAQUAR and ICQA allow multiple valid answers or object-level entity descriptions for a single image–question pair, whereas most widely used benchmarks, including VQA v2 and OK-VQA, implicitly contain answer variability but are frequently evaluated using a single-answer prediction paradigm.

Datasets such as Visual7W and VQA-Med provide limited answer variability compared to more recent structured benchmarks like KRVQA, which focuses on multi-step reasoning (that is crucial for the M2TVQA and ICQ-TransE model). Consequently, the choice of dataset reflects both the target application domain and the extent to which it supports reasoning over multiple visual entities.

In terms of answer multiplicity and ambiguity level, datasets such as VQA v2, OK-VQA, and ICQA provide multiple humans or programmatically generated answers per question, which introduces a high level of answer ambiguity and reflects the diversity of interpretations. In contrast, structured datasets such as GQA and RSVQA typically provide a single dominant answer derived from structured annotations or scene graphs, resulting in lower ambiguity but more deterministic reasoning tasks.

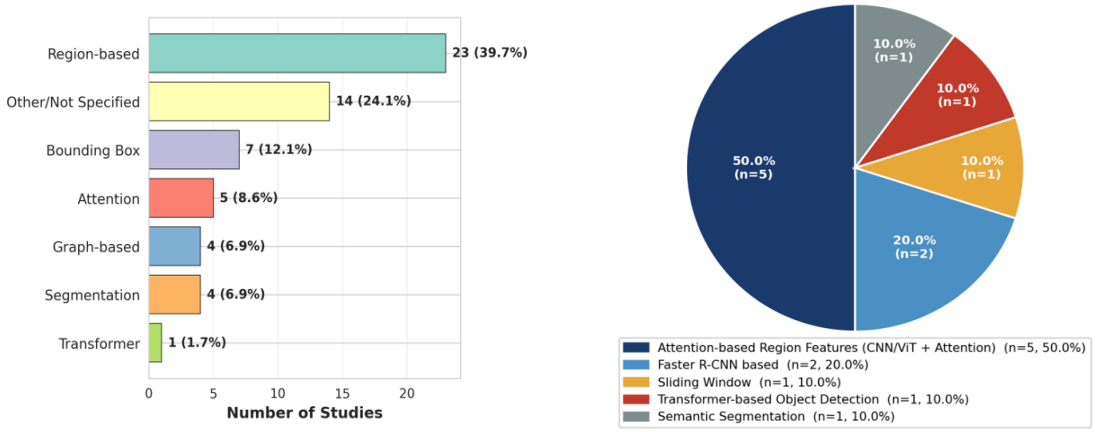
Table 7. Datasets employed across the explicit multi-answer VQA studies

Dataset	#Images	Avg. Answers	Annotation style	Ambiguity level	Object-level annotations	Multi-answer capability	Key characteristics
VQA v2.0	~204k	10	10 human answers	High	None	Implicit	Consensus-based evaluation
GQA	~113k	1	Scene graph	Low	Scene graphs	Implicit	Relational reasoning focus
OK-VQA	~14k	10	Multiple human answers	High	Region features	Implicit	Outside knowledge required
KRVQA	32,910	~4.8	Knowledge routed	Low	193,449 triplets	Implicit	Knowledge-routed reasoning
Visual7W	~47k	4	Multiple choice	Moderate	Bounding boxes (BBs)	Limited	Visual grounding focus
DAQUAR	~1.4k	1–3	Human annotations	Moderate	Object labels	Explicit	Object presence and counting
ICQA	~42k	Multiple	Programmatically generated	High	Icon object labels	Explicit	Specifically for multi-answer
RescueADI	4,044	~3.3	Structured tuples (P, R, A)	Moderate	16,949 masks, 14,483 BBs	Explicit	Disaster interpretation tuples
RSVQA	~1M	1	Automatically generated	Low	Bounding boxes	Implicit	Urban remote sensing data
Rad-ReStruct	3,720	~48	3-tier hierarchy	Moderate	178 pathol. vocab	Explicit	Hierarchical clinical reports
VQA-Med	~4k	1	Expert agenerated nnotations	Low-Moderate	None	Limited	Clinical QA benchmarks
HortiVQA-PP	14,460	6.3	Pest-predator matching	Moderate	Multi-label categories	Explicit	Pest-predator co-occurrence
Exp-VQA	~5.8M	Multi-scale	Multi-scale captioning	Moderate	9 facial regions	Explicit	Macro to micro-expressions

2.3.4. Object-based visual representations

These object-level representations serve as the primary visual inputs for subsequent reasoning and answer prediction, forming a consistent visual foundation across both general-purpose and domain-specific VQA studies as shown in Figure 4. While Figure 4(a) shows that region-based features are the most commonly adopted approach across all 58 studies (39.7%), Figure 4(b) reveals a different pattern among the 10 explicit multi-answer studies: attention-based region features emerge as the dominant strategy (50%, n=5), followed by Faster R-CNN-based detection (20%, n=2) and sliding-window detection, transformer-based

object detection, and semantic segmentation, each representing 10.0% (n=1). This shift suggests that researchers designing explicit multi-answer systems tend to favor attention mechanisms over general region-based approaches, likely because attention allows models to selectively focus on multiple visual regions simultaneously. This observation further highlights the importance of object-level representations as a fundamental component in enabling multi-entity reasoning in multi-answer VQA systems.



(a) (b)

Figure 4. A consistent visual foundation across both general-purpose and domain-specific VQA (a) object detection approaches (All Studies, n=58) and (b) distribution of visual representation methods among explicit multi-answer VQA models (n=10)

3.3. Narrative synthesis of the included studies

To synthesize the findings of the reviewed studies, the literature is analyzed according to the four analytical dimensions illustrated in Figure 5, namely task formulation, visual representation strategies, sources of answer multiplicity, and evaluation strategies.

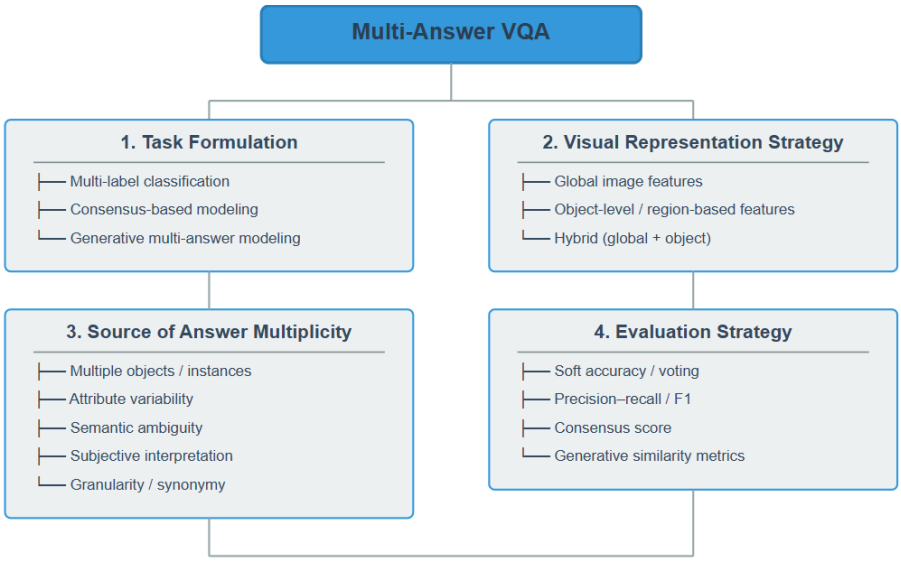


Figure 5. Taxonomy multi-answer VQA

3.3.1. Multi-answer vs. implicit multi-answer formulation (task formulation): RQ1

The first dimension of the taxonomy concerns how the VQA task is formally defined to accommodate multiple valid answers. Three main task formulations emerge from literature. The most explicit

approach models VQA as a multi-label classification problem, where multiple answers can be predicted simultaneously for a single image–question pair. This formulation is particularly prevalent in domain-specific applications, such as medical imaging and agriculture, where questions naturally correspond to multiple attributes, conditions, or findings.

A second approach relies on consensus-based modeling, in which multiple answers are derived from annotator agreement or probabilistic confidence distributions. This formulation is commonly associated with general-purpose VQA benchmarks that provide multiple human annotations per question, allowing systems to reflect answer variability without explicitly redefining the output space.

The third formulation adopts generative multi-answer modeling, where models are designed to produce a set or sequence of plausible answers rather than a single prediction. While this approach aligns well with recent advances in large vision–language and multimodal language models, it remains underexplored in the reviewed literature due to challenges in controllability and evaluation.

Research on multi-answer VQA demonstrates fundamental divergences in architectural strategies for handling answer multiplicity. Based on narrative synthesis, a critical appraisal of the 10 explicit studies (17%), as detailed in Table 6, reveals two distinct architectural cohorts:

a) Truly explicit models (n=3), which introduce architectural modifications designed to generate answer sets through mechanisms such as multi-label prediction or region aggregation. These models introduce direct architectural modifications to the output layer that enable the simultaneous generation of multiple answers [8], [24], [45].
b) Aggregation and loss innovations: SoftMaxUnion was introduced in the EMQA system to aggregate probabilities from image glimpses, generating answers simultaneously via spatial output map vectors [8]. SCAG employed sigmoid restriction loss or multi-target restriction loss to accommodate multiple ground-truth labels, yet inference returns the highest-scoring answer [70].
c) Structured and multi-label outputs: HortiVQA-PP and HiCA-VQA utilize sigmoid activation heads to predict multiple pest-predator relationships or pathological attributes concurrently [24], [45]. RescueADI produces structured output tuples (P, R, A) to record plans and intermediate results; however, the output element A remains a single answer in natural language [36].
d) Explicit Task Models (n=7): Upadhyay and Tripathy [49], GAT2R [63], ICQ-TransE [73], M2TVQA [74], Exp-VQA [78], SCAG [70], and RescueADI [44]. These models address multi-entity or complex task reasoning but remain constrained to single-answer outputs at inference (e.g., SCAG applies multi-label loss only during training, while RescueADI outputs structured task tuples with a single natural language answer).
e) Relational and scaled reasoning: graph attention networks (GAT) were used for explicit relational modeling [74], and Exp-VQA provides scale-specific captioning (macro, meso, micro) to describe facial actions in a single narrative response [49].
f) Refinement and triplets: a two-stage refinement of Top-4 candidate sets have been employed [73], while triplet reasoning to predict a specific tail entity based on multi-region selection has been adopted in knowledge-based VQA frameworks, M2TVQA and ICQ-TransE [63], [78].

This taxonomy confirms a significant gap in the field: while contemporary VQA models have developed robust perceptual capability of identifying and grounding multiple entities, supported by 83% (n=48) of the reviewed studies, the primary bottleneck remains in how they ‘speak’ (the output mechanism). Our analysis reveals that only 17% (n=10) of the studies explicitly formulate the task to address answer multiplicity.

3.3.2. Visual Representation Strategy: RQ2

The second dimension of this taxonomy is that visual representation strategies in multi-answer VQA have evolved from holistic-scene encoding to more granular, entity-centric representations. Based on the reviewed studies, these strategies can be broadly categorized into three groups: global image representations, object-level or region-based representations, and hybrid strategies that combine both.

Global image features capture holistic visual information from the entire scene. However, such representations are generally insufficient for multi-answer reasoning because they lack the spatial and instance-level granularity required to distinguish multiple relevant entities within complex images. As a result, object-level or region-based representations have emerged as the dominant strategy in multi-answer VQA. These representations are typically derived from object detection or region proposal mechanisms, enabling models to localize multiple objects and extract their attributes so that different visual entities can be associated with different plausible answers. Another emerging direction involves hybrid visual representations, which combine global contextual information with localized object-level features.

Based on an analysis of 10 explicit studies, region-based representation has become the dominant strategy, with several key architectural variations:

Detector technical comparison: region-based detectors such as Faster R-CNN remain widely used due to their semantically rich object features that support visual grounding in VQA systems, as demonstrated in models such as SCAG, ICQ-TransE, and M2TVQA. In contrast, YOLO-based detectors are commonly used in implicit studies because they prioritize computational efficiency and faster inference. As a result, YOLO is frequently applied in practical real-time scenarios such as robotics, autonomous systems, and agricultural monitoring [79], [80]. More recent architectures such as DETR (DEtection TRansformer) introduce transformer-based detection that eliminates the traditional region proposal stage (as in Faster R-CNN) and models relationships between objects directly through a global attention mechanism. GroundingDINO is one variant of the transformer-based detector in the explicit study used in the RescueADI framework.

Segmentation and hybrid representation: for higher precision, HortiVQA-PP integrates the segment anything model (SAM) to generate pixel-level region masks. Recent trends also show a shift toward hybrid representations using Vision Transformers (ViT) and foundation models (such as GIT, PubMedCLIP, and InstructBLIP), which naturally capture global context and local details simultaneously.

Structural and relational: structural reinforcement is achieved through graph-based relational reasoning. The GAT2R model uses GAT to model inter-object dependencies in a dynamic scene graph, which is crucial for generating a comprehensive answer set.

Overall, the distribution in Figure 4(b) confirms that while traditional region features provide the foundation (20%), there is a significant shift toward attention-based mechanisms (50%) and Transformer architectures. The flexibility of attention mechanisms proves to be more effective in mapping the various visual cues needed to support multi-answer reasoning in modern VQA systems.

3.3.3. Sources of answer multiplicity

The third dimension of taxonomy describes the main sources of answer multiplicity referenced in the literature. A synthesis of the studies examined identified four key factors. First, visual multiplicity arises when multiple objects in an image are relevant to the same question, as in ICQA and HortiVQA-PP. Second, attribute variability occurs when a single object has multiple valid attributes (e.g., an individual wearing a hat and glasses). Another example is medical VQA tasks, like Rad-ReStruct, which have more than one clinical attribute identified for a single visual finding. Third, semantic granularity produces answers at varying abstraction levels (e.g., “dog”, “animal”, “golden retriever”), explored in Exp-VQA. Fourth, subjective interpretation leads to divergent yet valid answers, especially when the task has to do with emotions or social situations. This becomes apparent in samples having several human annotations, such as VQA v2.0, where each question has multiple human annotations. These sources demonstrate that answer multiplicity is not simply annotation noise but reflects the inherent complexity of visual reasoning tasks.

3.3.4. Evaluation strategies in multi-answer VQA: RQ3

Evaluating multi-answer VQA systems presents distinct challenges. Traditional single-answer performance measures such as accuracy are insufficient for assessing the completeness of answer sets, and open-ended evaluation approaches often fail to capture the full range of semantically correct responses. Multiple correct answers have been identified as a major evaluation bottleneck in VQA tasks [7]. Recent research has explored alternative frameworks, including multi-label classification metrics, consensus-based scoring methods, and set-based similarity measures, to better handle predicted answer sets against multiple valid responses.

The fourth dimension highlights the lack of standardized evaluation protocols in multi-answer VQA. A key gap identified is the continued use of single-answer accuracy metrics in systems that implicitly support multiplicity, often leaving the potential for object-level reasoning undermeasured. As summarized in Table 8, evaluation strategies have evolved alongside output mechanisms, shifting from simple classification to set-based metrics (F1-score, Hamming Loss), linguistic metrics (BLEU, ROUGE, METEOR), and AI-based scoring (GPTScore) to better capture the diversity of semantically valid answers.

Table 8. Evaluation metrics in explicit multi-answer VQA

Strategy category	Main metric	Output mechanism	Related studies
Consensus-based	VQA accuracy (Soft accuracy)	SoftMax (Single-best)	GAT2R; ICQ-TransE
Set-based (multi-label)	F1-score, Precision, Recall	Sigmoid/SoftMaxUnion	Hosseinabad (2021); HiCA-VQA
Structural error	Hamming loss, Macro-F1	Sigmoid heads	HortiVQA-PP
Linguistic and semantic	BLEU, METEOR, ROUGE-L, Ada-similarity	Generative captioning	Exp-VQA
AI-augmented	GPTScore, GPT expert ratings	Autonomous Agent/LLM	RescueADI; HortiVQA-PP
Task-specific	Valid rate, mIoU, mAP	Structured Tuples	RescueADI

3.3.5. Conceptual framework of multi-answer VQA

Based on the synthesis of the reviewed studies, a conceptual framework for multi-answer VQA systems can be derived. The framework consists of several key stages including object detection for visual grounding, region-level feature extraction, multimodal fusion between image regions and question representations, and reasoning modules that associate visual entities with candidate answers. Finally, an answer generation mechanism produces either single or multiple responses depending on the task formulation. This framework highlights the central role of object-level visual representations in enabling multi-answer reasoning, as seen in Figure 6.

Multimodal fusion mechanisms play a central role in determining how visual and textual information interact during reasoning. Based on the reviewed studies, three common fusion strategies can be identified. Early fusion integrates visual and textual features at the input stage, allowing joint feature learning but often lacking fine-grained alignment between modalities. Late fusion combines independently processed visual and textual representations at the decision stage, which simplifies architecture but may limit cross-modal interaction. More recent approaches employ cross-modal attention or transformer-based fusion, enabling dynamic interactions between image regions and question tokens. Such attention-based fusion strategies are particularly advantageous for multi-answer VQA because they allow the model to associate different visual entities with multiple potential answer candidates.

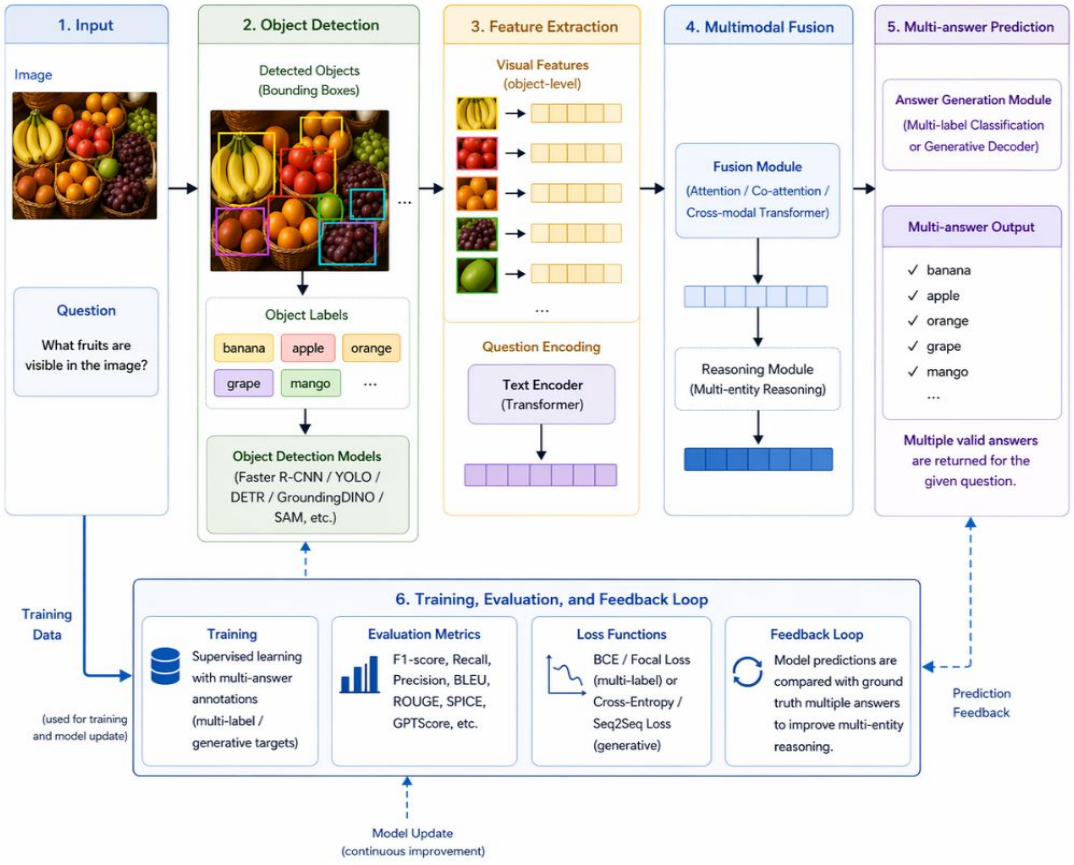


Figure 6. Conceptual framework of multi-answer VQA with object detection

3.4. Discussion

This study investigated the current state of multi-answer VQA with emphasis on object detection integration. While earlier studies have explored transformer-based architectures, graph-based reasoning, and domain-specific VQA applications [15]–[20], none have specifically examined how VQA systems accommodate multiple valid answers through object-level visual representations. This gap motivated this systematic review. By providing a structured synthesis of multi-answer VQA systems, this review contributes to the field and, to the best of our knowledge, is among the first studies to explicitly analyze the relationship between object-level visual grounding and answer multiplicity.

The findings collectively address the four research questions guiding this investigation. Addressing RQ1, the results indicate that multi-answer VQA remains an emerging paradigm rather than an established methodological standard, presenting both challenges and opportunities for the research community. This systematic review has documented that only 17% (n=10) of studies employing object-level visual representations explicitly formulate multi-answer generation as a primary objective, while the remaining 83% (n=48) of included studies implicitly support answer variability through enhanced visual grounding mechanisms without fundamentally reimagining the answer generation process. Three methodological approaches were identified: multi-label classification, consensus-based modeling, and generative multi-answer modeling.

Regarding RQ2, object detection contributed to multi-answer reasoning primarily through attention-based region features (50% of explicit studies) as in Figure 4(b) and region-based CNN features (39.7% of all studies) as in Figure 4(a). This shift suggests that researchers designing explicit multi-answer systems tend to favor attention mechanisms over general region-based approaches, likely because attention allows models to selectively focus on multiple visual regions simultaneously. Object-level representations such as bounding boxes, region proposals, instance masks, and region-based features enable models to identify and reason about multiple entities within complex scenes. These representations facilitate visual grounding, support disambiguation of ambiguous queries, and enable models to associate different visual entities with distinct answer candidates. However, the full potential of object detection for multi-answer reasoning remains largely unrealized due to insufficient integration between visual grounding mechanisms and answer generation modules. This observation is further supported by the study distribution identified in this review, where 83% of the included studies rely on implicit multi-answer reasoning rather than explicitly modeling multiple-answer outputs.

With respect to RQ3, the most employed datasets were VQA 2.0, GQA, and OK-VQA for general-purpose tasks and RSVQA and VQA-Med for domain-specific applications, while evaluation metrics ranged from soft accuracy and F1-score to Hamming Loss and semantic similarity. Notably, analysis of the 10 explicit studies reveals a marked shift toward evaluations aligned with the output mechanisms. As output mechanisms evolved, evaluation strategies shifted from simple classification to set-based metrics (F1-score, Hamming Loss), linguistic metrics (BLEU, ROUGE, METEOR), and AI-based scoring (GPTScore) to capture diverse valid answers.

Addressing RQ4, several key limitations remain in current multi-answer VQA research. The field shows substantial heterogeneity in task formulations, evaluation metrics, and methodologies, complicating cross-study comparison and limiting cumulative knowledge development. In addition, the limited availability of datasets explicitly designed for multi-answer VQA remains a major challenge, as most existing benchmarks were originally developed for single-answer prediction. This limitation is further compounded by the absence of standardized multi-answer benchmarks, inconsistent evaluation protocols, and a technical gap between task design and output architecture.

Compared to previous studies, these findings extend our understanding of the multi-answer VQA landscape in important ways. The dominance of implicit approaches (83%) aligns with [16] and [17], which noted that most VQA systems prioritize single-answer prediction despite having the architectural capacity for multi-object reasoning. However, this review further reveals a critical architectural bottleneck: a technical lag in model design. Specifically, out of 10 explicit studies (17%), sample classification shows a precise split between truly explicit models (n=3) and explicit task models (n=7). While explicit task models acknowledge answer multiplicity at the task level, they remain technically constrained by single-output architectures, is a unique barrier to multi-answer VQA not previously identified in existing surveys.

Furthermore, while evaluation inconsistency is a known hurdle in open VQA [7], this issue becomes far more critical in multi-answer contexts, where evaluating the completeness and recall of the predicted answer set adds another layer of complexity.

This review has several limitations. First, restriction to Scopus and journal articles only may have excluded relevant conference papers from CVPR, NeurIPS, and ICML. Second, substantial heterogeneity in task definitions and evaluation metrics precluded quantitative meta-analysis. Third, the broader operational definition of "implicit multi-answer" adopted here may affect comparability with future reviews that use stricter criteria.

These findings carry important implications. The technical lag identified in five of ten explicit studies suggests that architectural innovation, particularly in output layer design, is a more urgent priority than new dataset creation. The underrepresentation of agricultural VQA (n=2, 7.1% of domain-specific studies), despite its natural suitability for multi-answer tasks, represents a significant research opportunity. Detailed future directions are outlined in Section 3.5.

Overall, the results of this review demonstrate that while object detection has significantly improved visual grounding in VQA systems, the transition from implicit reasoning to explicit multi-answer modeling remains an open challenge for the research community.

3.5. Future research directions: RQ4

The results of this review show that there are several research directions that can be used to direct research efforts in the future in multi-answer VQA.

a) Explicit multi-answer formulation paradigms: future studies should move beyond implicit reasoning approaches and develop explicit formulations that treat answer multiplicity as a core characteristic of visual reasoning. This requires frameworks that regularly model sources of answer variability such as visual multiplicity, semantic ambiguity, contextual interpretation, and subjective perception.
b) Standardized evaluation frameworks: another area of research will be the creation of standardized evaluation frameworks. Protocols for evaluating multi-answer VQAs will require further research into set-based metrics, semantic similarity metrics, and consensus-based scoring strategies to obtain greater answer diversity with continued semantically correct answer(s).
c) Object detection integration architectures: continued advancement will require a closer connection between object detection modules and the answer-generation modules; improved attention mechanisms, relational reasoners, and structured output architectures could mean that the model digs in harder to the object-based information when reasoning about multiple answers.
d) Large vision-language models (VLMs) and multimodal LLMs: recent advances in large VLMs and MLLMs, such as BLIP-2 and Flamingo, offer promising alternatives for multi-answer VQA through diverse response generation. However, grounding AI-generated answers in detected visual entities remains challenging, as generating multiple answers may increase the risk of hallucinated outputs. Future research should focus on tighter integration between object detection modules and generative MLLMs to ensure multiple answers are both visually grounded and semantically relevant.
e) Dataset development and annotation protocols: relatively few datasets are currently available for multi-answer VQA research. Future multi-answer VQA datasets should have annotation protocols that help to record valid variations in acceptable answer forms while also providing clear annotation standards to avoid confusing inconsistent annotating with valid answer diversity.
f) Domain-specific applications and transfer learning: use of multi-answer VQA in specific domains (including medical imaging, agriculture, and remote sensing) holds great potential since they exhibit many attributes in relevant questions. Transfer learning approaches can be useful in making general models of VQA available for use in these specialty domain applications.
g) Interpretability and explainability mechanisms: interpretability is critical for high-stakes deployments such as medical diagnosis and autonomous systems. Future research should explore attention maps, region attribution strategies, and post-hoc techniques such as Grad-CAM and SHAP to elucidate the visual regions influencing each generated response, shifting the focus from raw accuracy to explainable multi-answer reasoning.
h) Human–computer interaction considerations: future systems should investigate effective presentation of multiple answers through ranked responses, confidence scores, and interactive visual explanations, while minimizing cognitive load and supporting user exploration of alternative answers in decision-support scenarios.

4. CONCLUSION

This SLR examined the emerging field of multi-answer VQA, with particular emphasis on the role of object detection in supporting answer multiplicity. By reviewing 58 peer-reviewed journal articles published between 2020 and 2025, this study provides a structured synthesis of methodological approaches, visual representation strategies, datasets, evaluation metrics, and future research directions related to multi-answer VQA. The findings indicate that most current VQA systems still rely on implicit multi-answer reasoning through object-level visual grounding rather than explicitly generating multiple valid answers. Object detection plays a central role in enabling multi-entity reasoning by allowing models to associate different image regions with distinct semantic interpretations. However, a substantial gap remains between advances in visual grounding and the development of output mechanisms capable of producing multiple semantically valid responses.

This review also identifies major challenges related to dataset availability, evaluation inconsistency, and the lack of standardized benchmarks for multi-answer VQA. Existing evaluation protocols remain limited in assessing the completeness, semantic diversity, and visual grounding of multiple generated answers. To address these limitations, this paper proposes a four-dimensional taxonomy covering task formulation, visual representation strategies, sources of answer multiplicity, and evaluation methods.

Overall, the results suggest that future research should focus on explicit multi-answer architectures, visually grounded generative VLM/MLLM systems, standardized evaluation frameworks, and explainable reasoning mechanisms to support more reliable and context-aware multi-answer VQA systems.

ACKNOWLEDGMENTS

The authors acknowledge IPB University, Bogor, Indonesia, and the State Islamic University (UIN) Syarif Hidayatullah Jakarta, Indonesia, for their continuous support for this research, particularly through the Directorate of Research and Community Engagement.

FUNDING INFORMATION

We want to convey our gratitude to the National Zakat Agency of Indonesia (BAZNAS) Republic of Indonesia, the Ministry of Religious Affairs of the Republic of Indonesia, and the Indonesia Endowment Fund for Education (LPDP), Ministry of Finance of the Republic of Indonesia, for providing scholarship support for the completion of this research.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Nidaul Hasanati	✓	✓	✓		✓	✓	✓	✓	✓	✓			✓	✓
Taufik Djatna	✓	✓		✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
Imas Sukaesih Sitanggang		✓		✓	✓		✓	✓		✓	✓	✓	✓	
Arif Imam Suroso	✓	✓	✓	✓	✓		✓			✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest regarding the publication of this paper.

DATA AVAILABILITY

The datasets analyzed in this study were obtained from previously published articles cited in the reference list, and detailed study characteristics and the completed PRISMA 2020 checklist are available in the Supplementary Material at https://doi.org/10.5281/zenodo.19995261.

REFERENCES

[1] S. Antol *et al.*, "VQA: Visual question answering," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, pp. 2425–2433, 2015, doi: 10.1109/ICCV.2015.279.

[2] K. Kafle and C. Kanan, "Visual question answering: Datasets, algorithms, and future challenges," *Computer Vision and Image Understanding*, vol. 163, pp. 3–20, 2017, doi: 10.1016/j.cviu.2017.06.005.

[3] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets," *Visual Computer*, vol. 38, no. 8, pp. 2939–2970, 2022, doi: 10.1007/s00371-021-02166-7.

[4] N. Bhattacharya, Q. Li, and D. Gurari, "Why does a visual question have different answers?," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-Octob. pp. 4270–4279, 2019, doi: 10.1109/ICCV.2019.00437.

[5] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? Dataset and methods for multilingual image question answering," *Advances in Neural Information Processing Systems*, vol. 2015-Janua, pp. 2296–2304, 2015.

[6] D. Gurari and K. Grauman, "CrowdVerge: Predicting if people will agree on the answer to a visual question," *Conference on Human Factors in Computing Systems - Proceedings*, vol. 2017-May. pp. 3511–3522, 2017, doi: 10.1145/3025453.3025781.

[7] M. Luo *et al.*, "'Just because you are right, doesn't mean I am wrong': Overcoming a bottleneck in the development and evaluation of open-ended visual question answering (VQA) tasks," *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*. pp. 2766–2771, 2021, doi: 10.18653/v1/2021.eacl-main.240.

[8] S. H. Hosseinabad, M. Safayani, and A. Mirzaei, "Multiple answers to a question: a new approach for visual question answering," *Visual Computer*, vol. 37, no. 1, pp. 119–131, 2021, doi: 10.1007/s00371-019-01786-4.

[9] M. Güllü and N. Barişçi, "Visual question answering for intelligent communication systems: a systematic review," *IEEE Access*, no. December 2025, pp. 1–1, 2026, doi: 10.1109/access.2026.3654676.

[10] P. Anderson *et al.*, "Bottom-Up and Top-Down attention for image captioning and visual question answering," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. pp. 6077–6086, 2018, doi: 10.1109/CVPR.2018.00636.

[11] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models," *International Conference on Machine Learning*, 2023.
[12] J. B. Alayrac *et al.*, "Flamingo: a visual language model for few-shot learning," in *Advances in Neural Information Processing Systems*, 2022, vol. 35, doi: 10.48550/arXiv.2204.14198.
[13] B. Liu, L. M. Zhan, L. Xu, L. Ma, Y. Yang, and X. M. Wu, "Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering," *Proceedings - International Symposium on Biomedical Imaging*, vol. 2021-April, pp. 1650–1654, 2021, doi: 10.1109/ISBI48211.2021.9434010.
[14] Y. Zhao, S. Wang, Q. Zeng, and W. Ni, "Informed-learning-guided visual question answering model of crop disease," *Plant Phenomics*, pp. 1–16, 2024, doi: 10.34133/plantphenomics.0277.
[15] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: a survey," *ACM Computing Surveys*, vol. 54, no. 10, 2022, doi: 10.1145/3505244.
[16] A. A. Yusuf, C. Feng, X. Mao, R. Ally Duma, M. S. Abood, and A. H. A. Chukkol, "Graph neural networks for visual question answering: a systematic review," *Multimedia Tools and Applications*, vol. 83, no. 18, pp. 55471–55508, 2024, doi: 10.1007/s11042-023-17594-x.
[17] J. Zhang, J. Teng, W. Ding, and Z. Wang, "Cross-modal heterogeneous graph reasoning network for visual question answering," *Neural Computing and Applications*, 2025, doi: 10.1007/s00521-025-11261-y.
[18] Z. Lin *et al.*, "Medical visual question answering: A survey," *Artificial Intelligence in Medicine*, vol. 143. p. 102611, 2023, doi: 10.1016/j.artmed.2023.102611.
[19] D. Ding, T. Yao, R. Luo, and X. Sun, "Visual question answering in robotic surgery: a comprehensive review," *IEEE Access*, vol. 13, pp. 9473–9484, 2025, doi: 10.1109/ACCESS.2024.3525145.
[20] S. S. Noor Mohamed and K. Srinivasan, "A comprehensive interpretation for medical VQA: Datasets, techniques, and challenges," *Journal of Intelligent and Fuzzy Systems*, vol. 44, no. 4, pp. 5803–5819, 2023, doi: 10.3233/JIFS-222569.
[21] M. J. Page *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *The BMJ*, vol. 372, 2021, doi: 10.1136/bmj.n71.
[22] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering - A systematic literature review," *Information and Software Technology*, vol. 51, no. 1, pp. 7–15, Jan. 2009, doi: 10.1016/j.infsof.2008.09.009.
[23] K. Zou *et al.*, "Uncertainty-aware medical diagnostic phrase identification and grounding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025, doi: 10.1109/TPAMI.2025.3596878.
[24] J. Zhang, B. Li, and S. Zhou, "Hierarchical modeling for medical visual question answering with cross-attention fusion," *Applied Sciences (Switzerland)*, vol. 15, no. 9, 2025, doi: 10.3390/app15094712.
[25] J. Wang, K. P. Seng, Y. Shen, L. M. Ang, and D. Huang, "Image to label to answer: an efficient framework for enhanced clinical applications in medical visual question answering," *Electronics (Switzerland)*, vol. 13, no. 12, 2024, doi: 10.3390/electronics13122273.
[26] C. Shu *et al.*, "MITER: Medical Image–TExt joint adaptive pretRaining with multi-level contrastive learning," *Expert Systems with Applications*, vol. 238, 2024, doi: 10.1016/j.eswa.2023.121526.
[27] F. Cong, S. Xu, L. Guo, and Y. Tian, "Anomaly matters: an anomaly-oriented model for medical visual question answering," *IEEE Transactions on Medical Imaging*, vol. 41, no. 11, pp. 3385–3397, 2022, doi: 10.1109/TMI.2022.3185113.
[28] X. Hu *et al.*, "Interpretable medical image visual question answering via multi-modal relationship graph learning," *Medical Image Analysis*, vol. 97, 2024, doi: 10.1016/j.media.2024.103279.
[29] X. Huang and H. Gong, "A dual-attention learning network with word and sentence embedding for medical visual question answering," *IEEE Transactions on Medical Imaging*, vol. 43, no. 2, pp. 832–845, 2024, doi: 10.1109/TMI.2023.3322868.
[30] W. Tian *et al.*, "MOSMOS: Multi-organ segmentation facilitated by medical report supervision," *Biomedical Signal Processing and Control*, vol. 106, 2025, doi: 10.1016/j.bspc.2025.107743.
[31] A. Lubna, S. Kalady, and A. Lijiya, "Visual question answering on blood smear images using convolutional block attention module powered object detection," *Visual Computer*, vol. 41, no. 1, pp. 739–757, 2024, doi: 10.1007/s00371-024-03359-6.
[32] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "RSVQA: Visual question answering for remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8555–8566, 2020, doi: 10.1109/TGRS.2020.2988782.
[33] R. Ou, Y. Hu, F. Zhang, J. Chen, and Y. Liu, "GeoPix: A multimodal large language model for pixel-level image understanding in remote sensing," *IEEE Geoscience and Remote Sensing Magazine*, vol. 13.0, no. 3, pp. 324.0-337.0, 2025, doi: 10.1109/MGRS.2025.3560293.
[34] W. Zhang, M. Cai, T. Zhang, Y. Zhuang, and X. Mao, "EarthGPT: a universal multimodal large language model for multisensor image comprehension in remote sensing domain," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, 2024, doi: 10.1109/TGRS.2024.3409624.
[35] L. Zhu, X. Su, J. Tang, Y. Hu, and Y. Wang, "View-based knowledge-augmented multimodal semantic understanding for optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, 2025, doi: 10.1109/TGRS.2025.3532349.
[36] M. Zhang, F. Chen, and B. Li, "Multistep question-driven visual question answering for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, 2023, doi: 10.1109/TGRS.2023.3312479.
[37] Y. Gao, Z. Bai, M. Zhou, B. Jia, P. Gao, and R. Zhu, "Adaptive conditional reasoning for remote sensing visual question answering," *Remote Sensing*, vol. 17.0, no. 8, 2025, doi: 10.3390/rs17081338.
[38] Z. Zhao, C. Zhou, Y. Zhang, C. Li, X. Ma, and J. Tang, "Text-guided coarse-to-fine fusion network for robust remote sensing visual question answering," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 230.0, pp. 1.0-17.0, 2025, doi: 10.1016/j.isprsjprs.2025.08.029.
[39] A. Saha and S. Maji, "Enhanced RSVQA insight through synergistic visual-linguistic attention models," *IEEE Geoscience and Remote Sensing Letters*, vol. 22.0, 2025, doi: 10.1109/LGRS.2025.3592253.
[40] Y. Zhou, R. Ding, X. Yang, X. Jiang, and X. Liu, "AirSpatialBot: A spatially aware aerial agent for fine-grained vehicle attribute recognition and retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63.0, 2025, doi: 10.1109/TGRS.2025.3570895.
[41] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. R. Murphy, "FloodNet: a high resolution aerial imagery dataset for post flood scene understanding," *IEEE Access*, vol. 9, pp. 89644–89654, 2021, doi: 10.1109/ACCESS.2021.3090981.
[42] P. Wang *et al.*, "RingMoGPT: a unified remote sensing foundation model for vision, language, and grounded tasks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, 2025, doi: 10.1109/TGRS.2024.3510833.

[43] Z. Yuan, L. Mou, Q. Wang, and X. X. Zhu, "From easy to hard: learning language-guided curriculum for visual question answering on remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022, doi: 10.1109/TGRS.2022.3173811.
[44] Z. Liu, D. Zhao, B. Yuan, and Z. Jiang, "RescueADI: adaptive disaster interpretation in remote sensing images with autonomous agents," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, 2025, doi: 10.1109/TGRS.2025.3532594.
[45] Z. Li *et al.*, "HortiVQA-PP: multitask framework for pest segmentation and visual question answering in horticulture," *Horticulturae*, vol. 11, no. 9, 2025, doi: 10.3390/horticulturae11091009.
[46] S. Tripathi, M. T. Nafis, I. Hussain, and A. K. J. Saudagar, "Multimodal fine-tuning of llms for robust document visual question answering," *IEEE Access*, 2025, doi: 10.1109/ACCESS.2025.3615201.
[47] S. Chakraborty, G. Harit, and S. Ghosh, "How well do MLLMs understand handwritten legal documents? A novel dataset for benchmarking," *International Journal on Document Analysis and Recognition*, 2025, doi: 10.1007/s10032-025-00548-4.
[48] X. Li, B. Wu, J. Song, L. Gao, P. Zeng, and C. Gan, "Text-instance graph: exploring the relational semantics for text-based visual question answering," *Pattern Recognition*, vol. 124, 2022, doi: 10.1016/j.patcog.2021.108455.
[49] Y. Yuan, J. Zeng, and S. Shiguang, "Exp-VQA: Fine-grained facial expression analysis via visual question answering," *Pattern Recognition*, vol. 168.0, 2025, doi: 10.1016/j.patcog.2025.111783.
[50] H. S. Asri and R. Safabakhsh, "Advanced visual and textual co-context aware attention network with dependent multimodal fusion block for visual question answering," *Multimedia Tools and Applications*, 2024, doi: 10.1007/s11042-024-18871-z.
[51] A. A. Liu, Z. Lu, N. Xu, W. Nie, and W. Li, "Multi-type decision fusion network for visual Q&A," *Image and Vision Computing*, vol. 115, 2021, doi: 10.1016/j.imavis.2021.104281.
[52] C. Wu, J. Lu, H. Li, J. Wu, H. Duan, and S. Yuan, "Compound-attention network with original feature injection for visual question and answering," *Signal, Image and Video Processing*, vol. 15, no. 8, pp. 1853–1861, 2021, doi: 10.1007/s11760-021-01932-3.
[53] L. Li, K. Du, M. Gu, F. Hu, and F. Lyu, "Region collaborative network for detection-based vision-language understanding," *Mathematics*, vol. 10, no. 17, 2022, doi: 10.3390/math10173110.
[54] H. Xia, R. Lan, H. Li, and S. Song, "ST-VQA: shrinkage transformer with accurate alignment for visual question answering," *Applied Intelligence*, vol. 53, no. 18, pp. 20967–20978, 2023, doi: 10.1007/s10489-023-04564-x.
[55] J. Wang, M. Zhu, Y. Li, H. Li, L. Yang, and W. L. Woo, "Detect2Interact: localizing object key field in visual question answering with LLMs," *IEEE Intelligent Systems*, vol. 39, no. 3, pp. 35–44, 2024, doi: 10.1109/MIS.2024.3384513.
[56] Y. Zhu, D. Chen, T. Jia, and S. Deng, "A lightweight transformer-based visual question answering network with weight-sharing hybrid attention," *Neurocomputing*, vol. 608, 2024, doi: 10.1016/j.neucom.2024.128460.
[57] J. Shi, D. Han, C. Chen, and X. Shen, "KTMN: Knowledge-driven two-stage modulation network for visual question answering," *Multimedia Systems*, vol. 30.0, no. 6, 2024, doi: 10.1007/s00530-024-01568-6.
[58] S. Zhang, Y. Chen, Y. Sun, F. Wang, H. Shi, and H. Wang, "LOIS: looking out of instance semantics for visual question answering," *IEEE Transactions on Multimedia*, vol. 26, pp. 6202–6214, 2024, doi: 10.1109/TMM.2023.3347093.
[59] M. Agrawal, A. S. Jalal, and H. Sharma, "Enhancing visual question answering with a two-way co-attention mechanism and integrated multimodal features," *Computational Intelligence*, vol. 40, no. 1, 2024, doi: 10.1111/coin.12624.
[60] Y.-T. Li, Y.-J. Lin, and H.-Y. Kao, "GViG: low-resource generative visual grounding using prompt-based language modeling for visual question answering," *International Journal of Data Science and Analytics*, vol. 20.0, no. 7, pp. 6489.0-6505.0, 2025, doi: 10.1007/s41060-025-00835-7.
[61] C. Wang, J. Yang, Y. Zhou, and X. Yue, "COOKIE: commonsense knowledge-guided mixture-of-experts framework for fine-grained visual question answering," *Information Sciences*, vol. 695, 2025, doi: 10.1016/j.ins.2024.121742.
[62] K. Su, H. Su, J. Li, and J. Zhu, "Toward accurate visual reasoning with dual-path neural module networks," *Frontiers in Robotics and AI*, vol. 7, 2020, doi: 10.3389/frobt.2020.00109.
[63] H. Liu *et al.*, "Multi-granularity feature interaction and multi-region selection based triplet visual question answering," *IEEE Transactions on Big Data*, 2024, doi: 10.1109/TBDATA.2024.3453750.
[64] T. Eiter, N. Higuera, J. Oetsch, and M. Pritz, "A neuro-symbolic asp pipeline for visual question answering," *Theory and Practice of Logic Programming*, vol. 22, no. 5, pp. 1–15, 2022, doi: 10.1017/S1471068422000229.
[65] B. Birmingham and A. Muscat, "Multi spatial relation detection in images," *Spatial Cognition and Computation*, vol. 22, no. 3–4, pp. 293–327, 2022, doi: 10.1080/13875868.2021.1957897.
[66] S. Zhou, D. Guo, J. Li, X. Yang, and M. Wang, "Exploring sparse spatial relation in graph inference for text-based VQA," *IEEE Transactions on Image Processing*, vol. 32, pp. 5060–5074, 2023, doi: 10.1109/TIP.2023.3310332.
[67] D. Guo, C. Xu, and D. Tao, "Bilinear graph networks for visual question answering," *IEEE Transactions on Neural Networks and Learning Systems*, 2021, doi: 10.1109/TNNLS.2021.3104937.
[68] D. Gao, R. Wang, S. Shan, and X. Chen, "CRIC: A VQA dataset for compositional reasoning on vision and commonsense," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5561–5578, 2023, doi: 10.1109/TPAMI.2022.3210780.
[69] J. Zhang, X. Liu, and Z. Wang, "Latent attention network with position perception for visual question answering," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–11, 2024, doi: 10.1109/TNNLS.2024.3377636.
[70] Z. Liu, K. Yang, Y. Weng, Z. He, X. Liu, and H. Gao, "SCAG: semantic co-occurring attention guided alignment for knowledge-based visual question answering," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 21.0, no. 7, 2025, doi: 10.1145/3734220.
[71] Q. Li, X. Tang, and Y. Jian, "Learning to reason on tree structures for knowledge-based visual question answering," *Sensors*, vol. 22, no. 4, 2022, doi: 10.3390/s22041575.
[72] M. Ma, T. Tohti, Y. Liang, Z. Zuo, and A. Hamdulla, "A focus fusion attention mechanism integrated with image captions for knowledge graph-based visual question answering," *Signal, Image and Video Processing*, vol. 18, no. 4, pp. 3471–3482, 2024, doi: 10.1007/s11760-024-03013-7.
[73] S. Upadhyay and S. S. Tripathy, "Enhancing VQA with emphasis-based soft attention," *Signal, Image and Video Processing*, vol. 19.0, no. 13, 2025, doi: 10.1007/s11760-025-04710-7.
[74] Y. Miao, W. Cheng, S. He, and H. Jiang, "Research on visual question answering based on GAT relational reasoning," *Neural Processing Letters*, vol. 54, no. 2, pp. 1435–1448, 2022, doi: 10.1007/s11063-021-10689-2.
[75] L. An-An *et al.*, "Multi-stage reasoning on introspecting and revising bias for visual question answering," *ACM Transactions on the Web*, vol. 18.0, no. 4, 2024, doi: 10.1145/3616399.
[76] G. KV, A. Nambiar, K. S. Srinivas, and A. Mittal, "Linguistically-aware attention for reducing the semantic gap in vision-language tasks," *Pattern Recognition*, vol. 112, 2021, doi: 10.1016/j.patcog.2020.107812.

[77] Y. J. Lin, C. S. Tseng, and H. Y. Kao, "Relation-aware image captioning with hybrid-attention for explainable visual question answering," *Journal of Information Science and Engineering*, vol. 40, no. 3, pp. 649–659, 2024, doi: 10.6688/JISE.202405_40(3).0014.
[78] H. Liu, B. Wang, X. Li, Y. Sun, Y. Hu, and B. Yin, "ICQ-TransE: LLM-enhanced image-caption-question translating embeddings for knowledge-based visual question answering," *IEEE Transactions on Artificial Intelligence*, 2025, doi: 10.1109/TAI.2025.3575553.
[79] F. A. Junior, "Real-time oil palm fruit grading system using smartphone and modified YOLOv4," *IEEE Access*, vol. 11, no. June, pp. 59758–59773, 2023, doi: 10.1109/ACCESS.2023.3285537.
[80] A. Qur, Y. Herdiyeni, W. A. Kusuma, A. H. Wigena, and S. Suhesti, "Detection and calculation of stomatal density using YOLOv5 : a study of high-yielding patchouli varieties," *Journal of Image and Graphics*, vol. 12, no. 3, pp. 228–238, 2024, doi: 10.18178/joig.12.3.228-238.

BIOGRAPHIES OF AUTHORS

Nidaul Hasanati was born in Jakarta, Indonesia, in 1979. She graduated with a bachelor's degree in informatics engineering in 2001 and a master's degree in business information systems in 2005 from Gunadarma University. She is currently pursuing a doctoral degree in Computer Science at School of Data Science, Mathematics and Informatics, IPB University, Bogor, Indonesia. Her research interests include intelligent information systems, artificial intelligence, data analytics, data mining, data science, data warehousing, and text mining. She is a lecturer and researcher in the Information Systems Department of Syarif Hidayatullah State Islamic University (UIN) Jakarta, Indonesia, where she has been serving since 2015. Previously, she was a lecturer in the Informatics Department at Al-Azhar University Indonesia from 2007 to 2015. She is a member of Indonesian Association of Informatics Experts (IAII) and IEEE. She can be contacted at email: nidaul.hasanati@uinjkt.ac.id, and nidaulhasanati@apps.ipb.ac.id.

Taufik Djatna is a senior professor of Industrial and System Engineering (ISE), Division of Agroindustrial Engineering and Technology, Faculty of Engineering and Technology, IPB University, Bogor, Indonesia Currently he is the Deputy Dean of Academic, Student, and Alumni Affairs of the Faculty of Engineering and Technology, IPB University, Bogor. He is the author of six books and more than 100 articles. His research interests include Skyline, T op-K, Query, Kansei Engineering, Emotion, Semantic Differential, Bitcoin, Ethereum, and the Internet of Things. He also has a patent for the mobile application of dynamic local culinary recommendations based on user preferences and locations. He received the JSPS Young Asian Best Presented Paper in 2014, 2018, and 2020; IEEE conferences in 2010. He is also an IEEE professional member, ACM professional member, member of The Institution of Engineers Indonesia (PII), and the Indonesian Agro-industrial Association. He can be contacted at email: taufikdjatna@apps.ipb.ac.id.

Imas Sukaesih Sitanggang is a senior professor in School of Data Science, Mathematics and Informatics, IPB University, Bogor, Indonesia, focusing on data mining and geographic information systems. Currently, she is working as the Head of the Postgraduate Program in Computer Science, School of Data Science, Mathematics and Informatics, IPB University, Bogor, Indonesia. She has conducted research in data mining and data warehousing, focusing on spatial datasets. Her current projects are 'Developing an Early Warning System for Forest and Peatland Fires in Sumatra and Kalimantan using a Spatio-Temporal Data Mining Approach' and 'Online Analytical Processing for Indonesian Agricultural Commodities'. She can be contacted at email: imas.sitanggang@apps.ipb.ac.id.

Arif Imam Suroso is a professor of business analytics at the IPB Business School, Bogor, Indonesia, with a cross-disciplinary educational background encompassing Agriculture, Library Science, Computer Science, and Agricultural Economics. His experience includes managing university-owned business units and active involvement in higher education management practices. He has a deep interest in business analytics, agricultural economics, and sustainable agribusiness. He believes that knowledge is meaningful when utilized to build positive change and shared sustainability. He currently serves as Chair of the IPB Business School Senate and Chief editor of Journal of Management and Agribusiness (JMA). He is open to collaboration, research, and discussions that have a real impact on society and the environment. He can be contacted at email: arifimamsuroso@apps.ipb.ac.id.