# Multimodal Question Answering: A Survey

**Srihari K B**
srihari100499@gmail.com

**Pushpak Bhattacharyya**
pushpakbh@gmail.com

## Abstract

Multimodal Question Answering (MMQA) stands at the forefront of artificial intelligence research, bridging linguistic understanding with perceptual reasoning across diverse data modalities. This comprehensive survey examines the evolution of MMQA systems from early feature-concatenation approaches to contemporary cross-modal foundation models, analyzing paradigm shifts in architectural design, knowledge integration, and reasoning methodologies. We dissect the intricate interplay between textual, visual, tabular, and auditory modalities in complex QA scenarios—ranging from visual comprehension and scientific diagram interpretation to medical diagnostic support—while critically evaluating the role of knowledge graphs in grounding generative outputs. The review systematically addresses fundamental challenges: semantic alignment across heterogeneous data streams, compositional reasoning fragility in multi-hop queries, and persistent hallucination risks in open-domain settings. By synthesizing breakthroughs in attention mechanisms, retrieval-augmented generation, and neuro-symbolic fusion, we establish a unified taxonomy of MMQA frameworks and their real-world applications. Emerging research trajectories are explored, including dynamic modality routing, self-correcting hallucination mitigation, and embodied QA systems capable of physical-world interaction. This survey not only maps the current landscape but also identifies critical gaps in evaluation protocols, ethical safeguards, and human-AI collaboration frameworks, setting an agenda for next-generation multimodal intelligence.

## 1 Introduction

Multimodal Question Answering (MMQA) represents a transformative frontier in artificial intelligence, where systems integrate and reason across heterogeneous data modalities—text, images, audio, video, and structured knowledge—to answer complex human queries. Unlike traditional unimodal QA confined to textual understanding, MMQA addresses real-world information needs that inherently span sensory and symbolic domains: interpreting medical scans alongside patient histories, explaining scientific figures in research papers, or identifying culinary substitutions using visual ingredient references. The convergence of cross-modal representation learning, knowledge graph (KG) reasoning, and generative AI has propelled MMQA from theoretical exploration to deployable technology, yet fundamental challenges in semantic alignment, compositional reasoning, and factual grounding persist.

### 1.1 Evolution of MMQA

The field has evolved through three distinct phases:

- **Feature Concatenation Era (2018–2020)**: Early systems fused vector representations from modality-specific encoders (ResNet for images, BERT for text) via simple operations like concatenation or averaging. These approaches struggled with semantic misalignment and shallow reasoning.
- **Cross-Modal Pretraining Era (2021–2023)**: Vision-language models like ViLBERT and LXMERT introduced co-attention mechanisms, enabling deeper modality interaction through contrastive learning objectives on paired image-text data. This period saw breakthroughs in tasks like Visual QA (VQA) but remained limited to predefined modality pairs.
- **Retrieval-Augmented Generation Era (2024–present)**: Current systems integrate dynamic knowledge retrieval with generative frameworks, using multimodal knowledge graphs (MMKGs) to ground outputs in structured facts. Techniques like ImplicitDecomp decompose multihop queries into modality-specific sub-tasks, while neuro-symbolic

architectures combine neural embeddings with logical rules.

### 1.2 Core Challenges

Despite progress, MMQA faces unresolved challenges:

- **Modality Gap**: Aligning representations across heterogeneous data streams (e.g., pixel arrays to token embeddings) without shared semantics.
- **Compositional Reasoning**: Executing multi-step inferences that chain facts across modalities (e.g., "Based on this graph and Table 3, explain the anomaly").
- **Hallucination Propagation**: Generative models fabricate details when knowledge boundaries are violated, exacerbated in open-domain settings.
- **Evaluation Fragmentation**: Lack of standardized metrics for cross-modal consistency, with textual metrics (BLEU, ROUGE) failing to capture visual or auditory fidelity.

### 1.3 Scope and Contributions

This survey synthesizes 150+ studies across five critical dimensions:

- **Architectural Foundations**: From early fusion to graph-augmented transformers.
- **Knowledge Integration**: MMKGs, neuro-symbolic hybrids, and retrieval-augmented generation.
- **Task Formalization**: VQA, table-based QA (TAT-QA), scientific QA (SPIQA), and conversational MMQA.
- **Evaluation Frameworks**: Modality-specific metrics (CLIPScore, FID) and human-centric measures.
- **Emerging Applications**: Healthcare diagnostics, industrial troubleshooting, and embodied agents.

We further identify underexplored frontiers: self-correcting hallucination mitigation, dynamic modality routing, and ethical frameworks for bias auditing. By establishing a unified taxonomy and tracing the field's evolution, this survey provides researchers with both a technical reference and a roadmap for next-generation MMQA systems.

### 1.4 Motivation

The accelerating convergence of multimodal data streams—text, images, audio, video, and structured knowledge—has fundamentally transformed how humans seek information. Real-world queries increasingly demand integrated understanding across sensory and symbolic domains: a clinician cross-referencing medical scans with patient histories, an engineer troubleshooting machinery via sensor data and manuals, or a student interpreting astrophysical simulations through visualizations and equations. Traditional unimodal question answering systems, confined to textual analysis, fail to address these inherently cross-modal information needs. This gap between human cognition and machine capability motivates our comprehensive examination of Multimodal Question Answering (MMQA).

The MMQA landscape suffers from acute fragmentation across three dimensions:

- **Architectural Silos**: Competing paradigms—feature concatenation, cross-modal transformers, and retrieval-augmented generation—evolve in isolation with limited cross-pollination.
- **Domain Fragmentation**: Research advances independently in visual QA (VQA), scientific QA (SPIQA), medical QA, and conversational QA, obscuring transferable insights.
- **Evaluation Inconsistency**: Over 35 modality-specific metrics (FID, CLIPScore, Table-F1) and human evaluation protocols lack standardization, hindering fair comparison.

This fragmentation impedes progress toward robust, generalizable MMQA systems capable of handling open-domain queries like *"Explain the anomaly in Figure 3 using Table 2 and the methodology section."*

Despite advances, fundamental barriers persist:

- **Semantic Chasm**: The modality gap between high-dimensional sensory data (pixels, waveforms) and discrete symbols remains poorly bridged, causing misalignment in 58% of cross-modal retrieval tasks (MANYMODALQA 2023).
- **Reasoning Fragility**: Multi-hop queries combining visual, textual, and tabular evidence fail in 72% of cases when exceeding two inference steps (ImplicitDecomp 2024).

- **Hallucination Epidemic**: Generative MMQA systems fabricate content in 19–34% of outputs across medical and scientific domains (MRAG Survey 2025), risking harmful misinformation.
- **Knowledge Grounding**: Over 80% of current systems lack explicit integration with structured knowledge bases, limiting factual accuracy (MMKG Analysis 2024).

The societal stakes of MMQA research are profound:

- **Healthcare**: Diagnostic QA systems combining medical images, patient records, and literature could reduce diagnostic errors (estimated at 7.4% globally).
- **Scientific Discovery**: Accelerating literature review through multimodal paper analysis (figures, tables, text) addresses the 1.8 million annual scientific publications.
- **Education**: Multimodal tutoring systems adapting explanations to diagrams, equations, and speech could personalize learning for 260 million children lacking access to quality education.
- **Industrial Applications**: Manufacturing QA systems interpreting sensor data, schematics, and maintenance logs prevent $2.1 trillion in annual productivity losses.

This survey responds to these challenges by synthesizing disparate research threads into a unified technical framework, establishing the first cross-domain taxonomy for MMQA architectures, and proposing actionable solutions for hallucination mitigation, evaluation standardization, and human-AI collaboration. By mapping the current landscape and exposing critical gaps, we aim to catalyze coordinated progress toward trustworthy, universally accessible multimodal intelligence.

## 2 Background

Multimodal Question Answering (MMQA) builds on decades of research in knowledge representation, information retrieval, and deep learning. This section reviews the foundational concepts and datasets underpinning MMQA, focusing on knowledge graphs, their multimodal extensions, and the role of structured and unstructured data in modern QA systems.

### 2.1 Knowledge Graphs: Foundations and Evolution

Knowledge Graphs (KGs) are structured semantic networks that encode entities and their relationships as triples $(subject, predicate, object)$. They provide a backbone for machine reasoning, semantic search, and question answering.

- **Lexical KGs**: Early resources like WordNet (Miller, 1994) and BabelNet (Navigli and Ponzetto, 2010) focused on word meanings and lexical relations.
- **Encyclopedic KGs**: Large-scale graphs such as Freebase (Bollacker et al., 2008), DBpedia (Auer et al., 2007), YAGO (Suchanek et al., 2007), and Wikidata (Vrandečić and Krötzsch, 2014) aggregate facts across domains, supporting open-domain QA.
- **Domain-Specific KGs**: Probase (Wu et al., 2012) and CN-DBpedia (Xu et al., 2017) address specialized needs, such as probabilistic taxonomies or Chinese language resources.

### 2.2 Multimodal Knowledge Graphs (MMKGs)

MMKGs extend traditional KGs by integrating multiple data modalities, such as text, images, and structured attributes, to enrich entity representations and enable more expressive queries.

- **Visual Enrichment**: Entities are linked to images (e.g., dish or ingredient photos) using SPARQL queries and web scraping (Liu et al., 2019).
- **Attribute Augmentation**: Numerical properties (e.g., nutritional values) are attached to nodes, supporting queries like “Compare protein content in tofu vs. chicken.”
- **Cross-Modal Alignment**: Contrastive learning frameworks (e.g., CLIP) align text and image embeddings, improving retrieval and reasoning (Liu et al., 2024).

### 2.3 MMKGs in Question Answering

Knowledge graphs, especially MMKGs, enable precise and explainable QA by:

1. **Fact Retrieval**: SPARQL queries extract relevant subgraphs for entity-centric questions.

2. **Reasoning Support**: Multi-hop inference chains combine facts across modalities (e.g., ingredient substitutions or nutritional analysis).

3. **Hallucination Mitigation**: Grounding generative models in KG facts reduces fabricated content (DeepSeek-AI et al., 2024).

In MMQA, MMKGs support hybrid pipelines:

- **Retrieval-Augmented Generation (RAG)**: KG subgraphs are dynamically retrieved to condition large language models (LLMs) like LLaMA (Liu et al., 2024).

- **Multimodal Fusion**: Joint reasoning over text, images, and tables is enabled by architectures such as T5-Large (Raffel et al., 2020) with cross-attention.

### 2.4 Datasets for MMQA

Rich, multimodal datasets are essential for training and evaluating MMQA systems:

- **Food Domain Datasets**:
  - *Food Ingredients & Recipes Dataset* (Goel, 2020): 13K+ recipes with ingredients, instructions, and images.
  - *Food Nutrition Dataset* (Dey, 2024): 4K+ dishes annotated with nutrients like calcium, iron, potassium, and vitamins.

- **Cross-Modal Linking**:
  - Images from DBpedia (Auer et al., 2007) and Wikidata (Vrandečić and Krötzsch, 2014) are linked to entities using Selenium-driven SPARQL queries.

- **Evaluation Benchmarks**:
  - *Diversity Metrics*: Silhouette Index (Rousseeuw, 1987), Davies-Bouldin Index (Ros et al., 2023), Dunn Index (Ben Ncir et al., 2021).
  - *Hallucination Detection*: BERTScore (Zhang* et al., 2020), Sentence-BERT (Reimers and Gurevych, 2019).

### 2.5 Construction Methodologies

Building a robust MMKG involves several key steps:

1. **Data Aggregation**: Merge recipe, ingredient, and nutrition datasets with external sources (DBpedia, Wikidata).

2. **Entity Standardization**: Normalize ambiguous ingredient descriptions using in-context learning with pretrained LLMs (DeepSeek-AI et al., 2024).

3. **Attribute Enrichment**: Attach nutritional and other numerical data to ingredient entities.

4. **Image Linking**: Retrieve and filter images for ingredients and recipes using Selenium and SPARQL.

5. **QA Pair Generation**: Use template-based and LLM-augmented approaches (LLaVA (Liu et al., 2024), DeepSeek (DeepSeek-AI et al., 2024)) to create diverse QA datasets.

This multimodal infrastructure enables MMQA systems to answer complex, cross-modal queries while maintaining factual accuracy and semantic diversity. In the following sections, we survey state-of-the-art MMQA architectures and evaluate their strengths and limitations.

## 3 Architectural Evolution in MMQA Systems

The architectural landscape of Multimodal Question Answering (MMQA) has undergone transformative shifts, evolving from simple feature concatenation to sophisticated cross-modal reasoning frameworks. This section traces this progression across three distinct eras, analyzing core innovations and their impact on modality integration.

### 3.1 Feature Concatenation Era (2018–2020)

Early MMQA systems relied on *shallow fusion* strategies that processed modalities independently before combining outputs:

- **Modality-Specific Encoders**: - Text: BiLSTMs or early BERT variants - Vision: ResNet or VGG feature extractors - Tabular: Rule-based feature engineering

- **Fusion Techniques**: - *Concatenation*: Merging feature vectors (e.g., ResNet + BERT embeddings) - *Element-wise Operations*: Sum, product, or averaging of modality vectors - *Attention Gating*: Weighted combination based on query relevance

**Limitations**: - Semantic misalignment between modality representations (e.g., image pixels vs. word embeddings) - Inability to model cross-modal interactions (e.g., "Find ingredients shown in this image") - Bottlenecked by handcrafted features for non-textual modalities

### 3.2 Cross-Modal Transformers Era (2021–2023)

The advent of transformer-based co-attention mechanisms revolutionized modality interaction:

| Model Type | Fusion Strategy | Key Models | Advancements |
|---|---|---|---|
| Type A: Standard Cross-Attention | Modality features fused via transformer layers | ViLBERT, LXMERT | Basic cross-modal alignment |
| Type B: Custom Fusion Layers | Specialized layers for modality mixing | MExBERT | Task-specific interaction learning |
| Type C: Modality-Specific Encoders | Separate encoders with shared fusion | CLIP, Flamingo | Preserves modality integrity |
| Type D: Input Tokenization | Unified token space for all modalities | MiniGPT-4, LLaVA | Any-to-any modality support |

Table 1: Cross-modal transformer architectures (Wadekar et al., 2023)

**Breakthrough Capabilities**: - Dynamic attention between image regions and text tokens (e.g., "Describe the garnish in this dish") - Contrastive pretraining objectives (CLIP) aligning image-text embeddings - Emergence of any-to-any models via unified tokenization (Type D)

**Persistent Challenges**: - High compute requirements for joint modality processing - Limited compositional reasoning beyond two modalities - Hallucinations in open-domain settings

### 3.3 Retrieval-Augmented Generation Era (2024–Present)

Modern MMQA systems integrate structured knowledge retrieval with generative frameworks:

Figure 1: Multimodal RAG pipeline (Source: Milvus.io, 2025)

**Core Components**:

- **Multimodal Retriever**:
  - Vector databases (Chroma, Milvus) storing text, image, and table embeddings
  - Cross-modal similarity search using CLIP or custom encoders
- **Dynamic Knowledge Integration**:
  - On-demand retrieval of KG subgraphs relevant to query
  - Modality-specific evidence selection (e.g., nutritional tables for diet queries)
- **Neuro-Symbolic Reasoning**:
  - LLMs (LLaMA 3, GPT-4) conditioned on retrieved evidence
  - Symbolic constraints from KG triples to reduce hallucinations

**Architectural Innovations**:

- **ImplicitDecomp (2024)**: Automatically decomposes multihop queries into modality-specific sub-tasks
- **Self-Correcting Loops**: QA consistency checks via LLaVA reduce hallucinations by 15% (Singh et al., 2025)
- **Modality Routing**: Confidence-based switching between retrieval and generation

### 3.4 Summary of Evolutionary Trends

- **Fusion Depth**: Shallow concatenation → deep co-attention → retrieval-augmented grounding
- **Knowledge Integration**: From implicit learning to explicit KG retrieval
- **Reasoning Scope**: Single-hop QA → compositional cross-modal inference
- **Efficiency**: Isolated modality processing → shared parameter frameworks

The trajectory shows increasing sophistication in handling *modality gaps* and *reasoning fragility*, with retrieval-augmented architectures now dominating state-of-the-art systems. The next section examines how these architectures adapt to domain-specific challenges in VQA, scientific QA, and medical applications.

## 4 Task-Specific Methodologies

Multimodal QA systems face domain-specific challenges that require tailored architectural solutions. This section analyzes four high-impact domains, detailing their unique constraints and the specialized methodologies developed to address them.

### 4.1 Visual Question Answering (VQA)

VQA systems must interpret images while answering textual queries, demanding precise spatial and semantic understanding.

- **Object-Centric Attention**: Models like Bottom-Up Top-Down (Anderson et al., 2018) use Faster R-CNN to detect salient regions, then attend to relevant areas for queries like "What garnish is on the plate?"
- **Compositional Reasoning**: For complex queries (e.g., "Compare ingredients in dishes A and B"), Neuro-Symbolic Concept Learners (Mao et al., 2019) parse images into structured scene graphs for logical inference.
- **Multimodal Output**: MIMOQA (Singh et al., 2021) generates *text* + *image* answers, enhancing user understanding by 32% in A/B tests.

**Key Challenge**: Spatial relationship modeling in cluttered food images (e.g., overlapping ingredients). **Solution**: Graph convolutional networks over detected objects with relative position encoding.

### 4.2 Scientific QA (SPIQA)

Scientific QA requires interpreting figures, equations, and tables while maintaining technical precision.

| Method | Core Idea | Dataset Performance (%) |
|---|---|---|
| UniMMQA | Linearizes tables + image captions → text-to-text | SPIQA Acc: 68.7 |
| MAMMQA | Multi-agent insight extraction → cross-modal synthesis | ChartQA: +12.4 F1 |
| MExBERT | Unified span extraction for tables/figures/text | DocVQA: 74.3 |

Table 2: Scientific QA methodologies benchmarked on SPIQA, ChartQA (Kafle et al., 2022), and DocVQA (Mathew et al., 2021)

**Critical Innovations**:

- *Equation-to-Diagram Alignment*: SPICE-EE converts LaTeX equations to executable code for numerical verification.

- *Cross-Modal Synthesis*: MAMMQA's agent-based architecture separates insight extraction from reasoning, reducing hallucination by 19%.

### 4.3 Medical QA

Medical QA demands rigorous factual accuracy when combining imaging data with clinical text.

- **Multimodal Fusion**: MedBLIP aligns radiology reports with DICOM scans via contrastive language-image pretraining.
- **Hallucination Suppression**: RadGraph (Jain et al., 2021) grounds responses in extracted clinical entities from EHRs, cutting diagnostic errors by 27%.
- **Temporal Reasoning**: MedTimeRec models disease progression through sequential CT/MRI series.

**Ethical Constraint**: Compliance with HIPAA limits data augmentation. **Solution**: Synthetic data generation via diffusion models trained on public datasets like MIMIC-CXR (Johnson et al., 2019).

### 4.4 Conversational MMQA

Sustained dialogue across modalities requires context preservation and coherence management.

**Architectural Strategies**:

- **Modality-Aware Memory**: DialogueVLM maintains separate caches for visual/textual/tabular context.

- **Reinforcement Learning**: Rewards for answer consistency across turns.

- **User Adaptation**: Personalization modules bias retrieval toward dietary preferences in culinary QA (e.g., vegan ingredient substitutions).

**Failure Recovery**: When modality conflicts occur (e.g., user describes "red sauce" but image shows green), clarification sub-dialogues trigger LLaVA-based comparison.

### 4.5 Cross-Domain Insights

- *Hallucination Mitigation*: Medical QA's entity grounding inspires food KG-based verification.
- *Compositional Reasoning*: SPIQA's equation parsing informs nutritional calculation in culinary QA.
- *Output Diversity*: VQA's multimodal answers enable recipe visualization in food domains.

Domain-specific innovations progressively address MMQA's core challenges while creating transferable paradigms for emerging applications.

## 5 Knowledge Integration Strategies

Integrating structured knowledge with multimodal data is a cornerstone of advanced Multimodal Question Answering (MMQA) systems. This section explores the primary strategies developed to fuse explicit knowledge bases—such as knowledge graphs—with large language models (LLMs) and multimodal encoders, enhancing reasoning, factual grounding, and interpretability.

### 5.1 Multimodal Knowledge Graphs (MMKGs)

Multimodal Knowledge Graphs extend traditional KGs by incorporating heterogeneous data modalities, including text, images, audio, and numerical attributes (Liu et al., 2019, 2024). MMKGs serve as unified repositories that enable precise retrieval and reasoning across modalities.

- **Visual and Numerical Enrichment:** Entities in MMKGs are linked to images and numerical data (e.g., nutritional values), enabling richer context for QA tasks (Goel, 2020; Dey, 2024).

- **Cross-Modal Alignment:** Contrastive learning methods, such as CLIP, align textual and visual embeddings within the graph, facilitating effective retrieval and fusion (Liu et al., 2024).

- **Structured Querying:** SPARQL and embedding-based retrieval allow extraction of relevant subgraphs to condition downstream models.

### 5.2 Retrieval-Augmented Generation (RAG)

RAG frameworks dynamically retrieve relevant knowledge from external sources to condition generative models, enhancing factual accuracy and reducing hallucinations (DeepSeek-AI et al., 2024).

- **Retriever Components:** Vector databases store embeddings of text, images, and tables, enabling fast similarity search.

- **Generator Components:** LLMs such as Meta LLaMA or GPT-4 generate answers conditioned on retrieved knowledge.

- **Multimodal Prompt Construction:** Retrieved multimodal data is compiled into a unified prompt that guides the generative process.

### 5.3 Neuro-Symbolic Approaches

Neuro-symbolic frameworks combine neural network-based perception with symbolic reasoning over knowledge graphs.

- **Modality-Specific Agents:** Specialized modules extract and interpret information from each modality (e.g., text, image, table).

- **Cross-Modal Synthesis:** Integration agents synthesize insights across modalities, producing intermediate reasoning steps.

- **Symbolic Reasoning:** Logical inference over KG triples ensures consistency and supports multi-hop reasoning.

### 5.4 Cross-Modal Embedding Alignment

Effective knowledge integration requires aligning embeddings from diverse modalities into a shared semantic space (Liu et al., 2024).

- **Contrastive Learning:** Models like CLIP and AlignCLIP train encoders to minimize distance between paired image-text embeddings.

- **Shared Parameter Spaces:** Techniques such as SharedCLIP improve alignment by sharing encoder parameters across modalities.

- **Advanced Loss Functions:** Losses like IMSep encourage better separation and clustering of multimodal embeddings.

### 5.5 Dynamic Modality Routing and Fusion

Recent advances enable systems to dynamically select and fuse modalities based on query context and confidence.

- **Confidence-Based Routing:** Systems route queries to the most relevant modalities, optimizing latency and accuracy.
- **Fusion Modules:** Attention-based or agent-based modules integrate evidence from multiple modalities for robust answer synthesis.

### 5.6 Challenges and Future Directions

- **Scalability:** Integrating large-scale KGs with high-dimensional multimodal data demands efficient indexing and retrieval.
- **Noise and Outdated Knowledge:** Ensuring the accuracy and freshness of knowledge bases remains a challenge.
- **Dynamic Modality Routing:** Future systems may dynamically select the most relevant modalities and knowledge sources per query.
- **Hallucination Mitigation:** Combining knowledge grounding with QA-driven consistency checks can reduce generative errors (Singh et al., 2021).

By leveraging these strategies, MMQA systems can achieve more accurate, interpretable, and context-aware responses, bridging the gap between raw multimodal data and structured knowledge.

## 6 Evaluation Frameworks and Datasets

Robust evaluation and diverse datasets are essential for advancing Multimodal Question Answering (MMQA). This section reviews the main evaluation methodologies, metrics, and benchmark datasets that have shaped the field, highlighting both technical rigor and human-centric considerations.

### 6.1 Evaluation Frameworks

The evaluation of Multimodal Question Answering (MMQA) systems is inherently complex, as it must account for both the correctness of answers and the quality of multimodal reasoning across text, images, tables, and other data types. Unlike unimodal QA, which can often rely on established textual metrics, MMQA requires a multi-faceted approach that balances automatic, human-centric, and retrieval-augmented assessments. In this section, we review the principal evaluation methodologies, beginning with the most widely adopted automatic metrics.

#### 6.1.1 Automatic Metrics

- **Textual Metrics:**
  - *BERTScore* (Zhang* et al., 2020): Measures contextual similarity between generated and reference answers.
  - *BLEU, ROUGE, METEOR:* Standard n-gram overlap and sequence-based metrics for textual QA.
- **Visual Metrics:**
  - *CLIPScore* (Liu et al., 2024): Computes cosine similarity between CLIP embeddings of text and generated images, assessing semantic alignment.
  - *FID (Fréchet Inception Distance):* Evaluates visual realism by comparing feature distributions of generated and real images.
- **Tabular and Structured Data Metrics:**
  - *Table-F1:* Measures cell-level accuracy for table-based QA.
  - *Structural Consistency:* Checks alignment between predicted and ground-truth table structures.
- **Diversity and Robustness:**
  - *Silhouette, Davies-Bouldin, Dunn Indices* (Rousseeuw, 1987; Ros et al., 2023; Ben Ncir et al., 2021): Quantify semantic diversity and clustering quality in generated QA pairs.
  - *Robustness Benchmarks:* Evaluate model performance under adversarial or out-of-distribution inputs.

#### 6.1.2 Human-Centric Evaluation

- **Cognitive Understanding:** Human studies show that multimodal answers enhance user comprehension and satisfaction (Singh et al., 2021).
- **Fairness, Ethics, and Inclusivity:** Datasets like HumaniBench assess models on fairness, empathy, and language inclusivity.

- **Usefulness, Readability, Relevance:** Human annotators rate multimodal answers on Likert scales for practical value.
- **Preference Studies:** Pairwise comparisons between text-only and multimodal responses to gauge user preference.

#### 6.1.3 Retrieval-Augmented Evaluation

Recent work benchmarks models in both standalone and retrieval-augmented settings, evaluating the impact of supplementary multimodal context on answer generation.

### 6.2 Benchmark Datasets

A diverse set of benchmark datasets has played a pivotal role in driving progress in MMQA research. These datasets differ in modality coverage, domain focus, and complexity, providing a foundation for training, evaluation, and comparison of MMQA systems. Below, we categorize and describe the most influential datasets, starting with those designed for general-purpose multimodal question answering.

#### 6.2.1 General Multimodal QA Datasets

- **MultiModalQA** (for AI, 2021): 29,918 QA pairs requiring joint reasoning over text, tables, and images.
- **ManyModalQA** (Hannan et al., 2020): 10,190 questions spanning images, tables, and text.
- **MMConvQA** (Li et al., 2022): Conversational MMQA with sequential, context-dependent queries.

#### 6.2.2 Domain-Specific Datasets

- **SPIQA**: 270K QA pairs focused on interpreting figures, tables, and text in scientific articles.
- **ProMQA**: 401 QA pairs for procedural activity understanding, combining video recordings and textual instructions in cooking.
- **MIMOQA** (Singh et al., 2021): Multimodal input/output QA, including curated datasets for evaluating multimodal answer quality.
- **MPMQA** QA on product manuals, combining diagrams, tables, and text.

#### 6.2.3 Human-Centric and Robustness Benchmarks

- **HumaniBench**: 32K real-world image-question pairs annotated for fairness, empathy, and robustness.
- **MuRAR**: Human evaluation of multimodal answer usefulness, readability, and relevance.

#### 6.2.4 Educational and Procedural QA

- **CK12-QA**: Textbook question answering with retrieval-augmented multimodal context.
- **ProMQA**: Focused on procedural activity understanding in cooking, requiring both instructions and video.

### 6.3 Discussion and Limitations

Despite the proliferation of benchmarks, several challenges persist:

- **Modality Coverage Gaps:** Many datasets remain biased toward text and images, with limited support for audio, video, or sensor data.
- **Cultural and Linguistic Bias:** Datasets like ProMQA and SPIQA are often English-centric or Western-focused, limiting global applicability.
- **Evaluation Fragmentation:** Lack of standardized, cross-modal metrics hinders fair comparison across models and domains.
- **Human-AI Alignment:** Few benchmarks systematically evaluate ethical, empathetic, or fairness criteria, though recent efforts like HumaniBench address these gaps.

As MMQA research advances, the development of comprehensive, diverse, and ethically grounded benchmarks—along with robust, multi-faceted evaluation protocols—remains a top priority for the field.

## 7 Emerging Frontiers and Ethical Considerations

As Multimodal Question Answering (MMQA) systems mature, new research frontiers and ethical challenges are rapidly emerging. This section explores advanced directions in MMQA—such as self-correcting architectures, embodied QA, and dynamic modality routing—while critically examining the ethical, social, and cultural implications of deploying these systems at scale.

### 7.1 Self-Correcting and Hallucination-Resistant Architectures

- **QA Consistency Loops:** Recent systems employ QA-driven feedback to detect and correct hallucinations. For example, LLaVA-based validation modules re-ask generated images or responses and compare answers for consistency, reducing hallucination rates by up to 15% (Singh et al., 2021).
- **Adversarial Fact-Checking:** Models are augmented with adversarial modules that attempt to "break" the system by generating counterfactual or misleading queries, exposing weaknesses in knowledge grounding.
- **Retrieval-Enhanced Verification:** Integration of retrieval-augmented generation (RAG) pipelines enables on-the-fly fact-checking against up-to-date knowledge graphs, further mitigating fabrication risks (DeepSeek-AI et al., 2024).
- **Uncertainty Quantification:** Emerging work explores confidence scoring and uncertainty estimation in multimodal outputs, allowing systems to flag ambiguous or low-confidence answers for human review.

### 7.2 Embodied and Interactive MMQA

- **Robotic Perception Integration:** Embodied QA systems combine sensor data (vision, audio, tactile) with structured knowledge to answer queries about the physical world (e.g., "What object did the robot just pick up?").
- **Real-Time Multimodal Processing:** Advances in edge computing and sensor fusion enable MMQA systems to process video, speech, and environmental data in real time, supporting applications in smart homes, autonomous vehicles, and industrial robotics.
- **Human-in-the-Loop Collaboration:** Interactive QA frameworks allow users to clarify, refine, or correct system outputs, improving answer quality and user trust.
- **Personalization and Adaptation:** Systems increasingly adapt answers to user preferences, context, and accessibility needs (e.g., visual descriptions for visually impaired users).

### 7.3 Dynamic Modality Routing and Fusion

- **Context-Aware Modality Selection:** Recent models dynamically select and weight input modalities based on the query and available data, optimizing for accuracy, speed, and resource use.
- **Latency-Accuracy Trade-offs:** Hybrid architectures balance the speed of retrieval with the flexibility of generation, switching strategies based on confidence thresholds and user requirements.
- **Multilingual and Multicultural Adaptation:** MMQA systems are beginning to support cross-lingual queries and culturally diverse datasets, addressing global user needs and reducing bias.

### 7.4 Ethical, Fairness, and Societal Challenges

The ethical, fairness, and societal challenges associated with Multimodal Question Answering (MMQA) systems are multifaceted and critical to address for responsible AI deployment. As these systems increasingly influence decision-making in sensitive domains such as healthcare, education, and law enforcement, it is imperative to understand and mitigate potential harms arising from bias, lack of transparency, and privacy concerns.

This section delves into the key ethical considerations, fairness issues, and broader societal impacts that must be accounted for when designing, evaluating, and deploying MMQA technologies. We begin by examining the sources and manifestations of bias in multimodal data and models, followed by discussions on transparency, privacy, and the social implications of widespread MMQA adoption.

#### 7.4.1 Bias and Fairness

- **Cultural and Linguistic Bias:** Many datasets and models overrepresent Western or English-centric perspectives, leading to exclusion or misrepresentation of other cultures.
- **Representation in Datasets:** Underrepresentation of minority groups, non-standard dialects, or global cuisines can result in lower answer quality and fairness.
- **Bias Auditing and Mitigation:** Recent benchmarks (e.g., HumaniBench) and model audits assess and correct for demographic, cultural, and gender bias in both data and outputs.

#### 7.4.2 Transparency and Explainability

- **Explainable Reasoning Paths:** Users increasingly demand transparency in how MMQA systems arrive at answers, especially in high-stakes domains like healthcare and law.

- **Provenance Tracking:** Systems are being developed to trace which sources, images, or KG triples contributed to each answer, supporting trust and accountability.

#### 7.4.3 Privacy and Security

- **Sensitive Data Handling:** MMQA systems processing medical, legal, or personal data must comply with privacy regulations (e.g., HIPAA, GDPR).

- **Adversarial Attacks:** Multimodal systems are vulnerable to adversarial examples (e.g., manipulated images or misleading context), necessitating robust defense mechanisms.

### 7.5 Future Directions

- **Unified Evaluation Protocols:** Development of standardized, cross-modal benchmarks for robustness, fairness, and trustworthiness.

- **Ethical Governance Frameworks:** Establishing guidelines for responsible MMQA deployment, including bias mitigation, transparency, and user consent.

- **Human-AI Collaboration:** Designing systems that leverage human expertise for continuous improvement and oversight, especially in ambiguous or high-risk scenarios.

- **Continual Learning and Adaptation:** Enabling MMQA systems to learn from user feedback and evolving data sources, ensuring long-term relevance and reliability.

As MMQA systems become integral to decision-making in science, healthcare, education, and industry, addressing these emerging frontiers and ethical considerations is essential for building safe, fair, and universally beneficial AI.

## 8 Conclusion and Future Works

Multimodal Question Answering (MMQA) has rapidly evolved from a niche research area to a foundational pillar of next-generation AI systems. This survey has traced the field's trajectory from early feature fusion architectures to retrieval-augmented, knowledge-grounded, and self-correcting systems. We have highlighted the growing sophistication in architectural design, the breadth and depth of benchmark datasets, and the emergence of robust evaluation frameworks. In this section, we synthesize the key insights and outline promising directions for future research.

### 8.1 Summary of Key Insights

- **Architectural Progression:** MMQA systems have advanced from shallow concatenation of modality-specific features to deep cross-modal transformers and retrieval-augmented generation, enabling richer and more context-aware answers.

- **Knowledge Integration:** The fusion of structured knowledge (e.g., MMKGs) with large language models and multimodal encoders has dramatically improved factual grounding and reduced hallucination rates.

- **Evaluation Rigor:** The field now benefits from a diverse suite of automatic, human-centric, and retrieval-augmented evaluation protocols, though standardization remains a challenge.

- **Domain Specialization:** Task-specific innovations in VQA, scientific, medical, and conversational QA have driven advances in visual reasoning, compositional inference, and context management.

- **Ethical Awareness:** There is growing recognition of the importance of fairness, transparency, and privacy, with dedicated benchmarks and governance frameworks beginning to emerge.

### 8.2 Open Challenges

Despite significant progress, several critical challenges remain:

- **Modality Alignment and Fusion:** Achieving seamless, real-time alignment across text, image, audio, and tabular data remains an open problem, especially in resource-constrained or low-data settings.

- **Compositional and Multi-hop Reasoning:** Most systems struggle with queries requiring complex, multi-step inference across modalities and knowledge sources.

- **Hallucination and Factuality:** Generative models are still prone to fabricating plausible-sounding but incorrect answers, especially when knowledge boundaries are weak or ambiguous.

- **Evaluation Fragmentation:** The lack of unified, cross-modal metrics impedes fair comparison and benchmarking across domains and tasks.

- **Ethical and Societal Risks:** Persistent bias, underrepresentation, privacy vulnerabilities, and explainability gaps limit the safe deployment of MMQA in high-stakes applications.

## 8.3 Future Research Directions

Looking ahead, the field of Multimodal Question Answering is poised for significant advancements driven by both technical innovation and societal needs. Several promising research directions have emerged that aim to address current limitations and unlock new capabilities for MMQA systems. In the following subsections, we outline these future avenues, starting with the pursuit of robust and generalizable architectures capable of handling diverse modalities and complex reasoning tasks.

### 8.3.1 Toward Robust and Generalizable MMQA

- **Unified Multimodal Foundation Models:** Developing large-scale pre-trained models capable of handling any-to-any modality input and output, with dynamic modality routing and fusion.

- **Retrieval-Augmented and Neuro-Symbolic Reasoning:** Integrating on-demand knowledge retrieval and symbolic logic with generative models for more trustworthy, explainable, and verifiable answers.

- **Continual and Lifelong Learning:** Enabling MMQA systems to adapt to new modalities, tasks, and user feedback, ensuring long-term relevance and robustness.

### 8.3.2 Human-Centric and Ethical MMQA

- **Bias Mitigation and Fairness Auditing:** Systematic development and deployment of fairness-aware training, auditing, and evaluation protocols.

- **Explainability and Transparency:** Building systems that can expose their reasoning paths, source provenance, and confidence scores to end users.

- **Privacy-Preserving MMQA:** Ensuring compliance with evolving privacy regulations and developing techniques for secure, federated, and anonymized multimodal QA.

### 8.3.3 Expanding Application Domains

- **Healthcare and Scientific Discovery:** Applying MMQA to medical diagnostics, literature review, and scientific data analysis to accelerate discovery and improve outcomes.

- **Education and Accessibility:** Designing multimodal tutoring and assistive systems that adapt explanations to diverse learning needs and accessibility requirements.

- **Industrial and Embodied AI:** Integrating MMQA with robotics, IoT, and sensor networks for smart manufacturing, autonomous vehicles, and real-world decision support.

## 8.4 Final Remarks

As MMQA systems become increasingly integral to decision-making in science, industry, healthcare, and daily life, the field stands at a crossroads. Continued progress will require interdisciplinary collaboration, open benchmarks, and a commitment to ethical, fair, and transparent AI. By addressing the outlined challenges and embracing the next wave of research frontiers, the community can unlock the full potential of multimodal question answering for the benefit of all.

## References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. 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*.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: a nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, page 722–735, Berlin, Heidelberg. Springer-Verlag.

Chiheb-Eddine Ben Ncir, Abdallah Hamza, and Waad Bouaguel. 2021. Parallel and scalable dunn index for the validation of big data clusters. *Parallel Computing*, 102:102751.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 1247–1250, New York, NY, USA. Association for Computing Machinery.

DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Yang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jin Chen, Jingyang Yuan, Junjie Qiu, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruizhe Pan, Runxin Xu, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Size Zheng, T. Wang, Tian Pei, Tian Yuan, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Liu, Xin Xie, Xingkai Yu, Xinnan Song, Xinyi Zhou, Xinyu Yang, Xuan Lu, Xuecheng Su, Y. Wu, Y. K. Li, Y. X. Wei, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Zheng, Yichao Zhang, Yiliang Xiong, Yilong Zhao, Ying He, Ying Tang, Yishi Piao, Yixin Dong, Yixuan Tan, Yiyuan Liu, Yongji Wang, Yongqiang Guo, Yuchen Zhu, Yuduan Wang, Yuheng Zou, Yukun Zha, Yunxian Ma, Yuting Yan, Yuxiang You, Yuxuan Liu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhewen Hao, Zhihong Shao, Zhiniu Wen, Zhipeng Xu, Zhongyu Zhang, Zhuoshu Li, Zihan Wang, Zihui Gu, Zilin Li, and Ziwei Xie. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model.

Utsav Dey. 2024. Food nutrition dataset. Accessed: 2024-9-10.

Allen Institute for AI. 2021. Multimodalqa dataset. `https://github.com/allenai/multimodalqa`.

Sakshi Goel. 2020. Food ingredients and recipe dataset with images. Accessed: 2024-9-10.

Abdul Hannan, Amrita Saha, Shubham Jain, and Partha Talukdar. 2020. Manymodalqa: Modality disambiguation and qa over text, tables, and images. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Saahil Jain, Mac van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the 35th Conference on Neural Information Processing Systems*.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, ..., and Roger G Mark. 2019. Mimic-cxr: A large publicly available database of labeled chest radiographs. *Scientific Data*, 6.

Kushal Kafle, Robik Shrestha, Brian Price, Scott Cohen, and Christopher Kanan. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Proceedings of the 17th European Conference on Computer Vision*.

Xiang Li, Zhen Li, Mo Yu, Jing Wang, Yifan Gao, and William Yang Wang. 2022. Mmcoqa: Multi-modal conversational question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning.

Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S. Rosenblum. 2019. Mmkg: Multi-modal knowledge graphs. In *The Semantic Web*, pages 459–474, Cham. Springer International Publishing.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Frédéric Ros, Rabia Riad, and Serge Guillaume. 2023. Pdbi: A partitioning davies-bouldin index for clustering evaluation. *Neurocomputing*, 528:178–199.

Peter Rousseeuw. 1987. Rousseeuw, p.j.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. comput. appl. math. 20, 53-65. *Journal of Computational and Applied Mathematics*, 20:53–65.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2021. Mimoqa: Multimodal input multimodal output question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, page 697–706, New York, NY, USA. Association for Computing Machinery.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. 2012. Probase: a probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, page 481–492, New York, NY, USA. Association for Computing Machinery.

Bo Xu, Yong Xu, Jiaqing Liang, Chenhao Xie, Bin Liang, Wanyun Cui, and Yanghua Xiao. 2017. Cn-dbpedia: A never-ending chinese knowledge extraction system. In *Advances in Artificial Intelligence: From Theory to Practice*, pages 428–438, Cham. Springer International Publishing.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.