International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I11P115

Volume 74 Issue 11, 202-213, November 2026
© 2026 Seventh Sense Research Group®

Original Article

KEAF-Net: Knowledge-Enhanced Adaptive Fusion Network for Visual Question Answering with Multi-Hop Graph Reasoning

Noorbhasha Junnu Babu[1], S. P. Rajamohana[2]

[1,2]*Department of Computer Science, Pondicherry University, Karaikal Campus, Puducherry, India.*

[1]*Corresponding Author: junnubabu@pondiuni.ac.in*

Received: 19 June 2026 Revised: 11 September 2026 Accepted: 18 September 2026 Published: 25 November 2026

Abstract *- Knowledge-intensive Visual Question Answering (VQA) becomes difficult when the required answer depends on information that is not directly visible in the image. Recognizing depicted objects may be sufficient for ordinary questions, whereas knowledge-oriented questions additionally require facts about use, location, category, or commonsense relations. External knowledge graphs can supply such information, but broad retrieval also introduces distracting triplets. KEAF-Net addresses this problem in three stages. The Adaptive Knowledge Filter (AKF) evaluates candidate triplets against the joint image-question context and learns from sampled leave-one-out changes in prediction loss rather than answer-string overlap. The retained evidence is fused with visual regions and question tokens in a typed heterogeneous graph, after which Multi-Hop Semantic Reasoning (MHSR) updates the query state through GRU-based attention hops. Across three runs, KEAF-Net reports 68.7% accuracy on the OK-VQA evaluation split and 62.4% on the A-OKVQA direct-answer test split, with approximately 252M inference parameters. Removing AKF, HGAF, and MHSR lowers OK-VQA accuracy by 3.2, 3.8, and 1.9 points, respectively. These controlled ablations indicate that evidence screening before typed multimodal fusion, followed by short iterative reasoning, accounts for the measured improvements within the KEAF-Net configuration.*

Keywords *- Counterfactual supervision, External evidence selection, Gated recurrent inference, Knowledge graph retrieval, Typed multimodal representation.*

1. Introduction

In VQA, a model receives an image together with a natural-language query and returns a short answer [1, 2]. The difficulty is not limited to identifying visible objects. Some questions also require regional grounding, interpretation of the wording, and a fact that is not encoded in the pixels. A question about sports equipment may depend on the associated game; a landmark question may require geographic knowledge; and an everyday-object question may require knowledge of typical use. Pre-trained vision-language models have improved general VQA performance [3-5], but OK-VQA [6] and A-OKVQA [7] are designed to expose cases in which visual evidence alone is insufficient. These benchmarks are therefore useful for testing whether a model can connect what it sees with relevant external knowledge [8].

External knowledge enters KB-VQA mainly through structured retrieval or through the implicit knowledge stored in large language models. Structured approaches query resources such as ConceptNet [9], CSKG [10], or task-specific knowledge bases [11-13], whereas LLM-based approaches obtain background facts or reasoning traces from large pre-trained models [14-16]. Neither route guarantees that the returned evidence is useful for the current sample. A triplet may share an entity with the image and still be irrelevant to the question, while an LLM-produced fact can appear plausible without adequate visual support. Previous studies report that poorly controlled knowledge can reduce answer accuracy [13, 17]. This risk is amplified in graph-based systems such as KRISP [18] and MuKEA [19], because an irrelevant node can influence other nodes during message passing.

A specific research gap remains at the interface of knowledge selection, multimodal fusion, and iterative reasoning. Representative recent systems address only parts of this chain: MGRN [43] performs multi-hop reasoning over a graph selected by a Top-K strategy, CEIK [12] coordinates explicit and implicit knowledge through gating, and RK-VQA [13] uses rational knowledge-aware fusion in a decoder. These designs do not jointly provide a prediction-loss-based relevance signal for retrieved facts, preserve visual, textual, and knowledge elements as typed nodes during fusion, and then re-query that fused graph through a controlled multi-hop


This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

state update. Consequently, irrelevant evidence can enter message passing, modality-specific structure can be weakened during fusion, or additional reasoning depth can be introduced without an explicit control study. KEAF-Net targets this combined gap by screening candidate evidence before graph construction, retaining modality types during heterogeneous fusion, and selecting reasoning depth empirically.

KEAF-Net couples these three decisions within one answer-prediction pipeline. First, the Adaptive Knowledge Filter (AKF) uses sampled leave-one-out loss changes to supervise triplet selection without using answer-string overlap as the relevance target. Second, Heterogeneous Graph Adaptive Fusion (HGAF) keeps visual regions, question tokens, and retained knowledge as separate node types while constructing connectivity from five edge categories. Third, Multi-Hop Semantic Reasoning (MHSR) re-queries the fused graph through GRU updates; the depth experiment selects T = 3. The three components are optimized with the answer objective so that evidence admission, cross-modal interaction, and iterative reasoning serve the same prediction task.

Evaluation is reported on VQA v2.0, OK-VQA, A-OKVQA, and GQA. KEAF-Net reaches 68.7% on the OK-VQA evaluation split and 62.4% on the A-OKVQA direct-answer test split, using approximately 252M inference parameters. In the module ablations, HGAF causes the largest single performance loss when removed. Removing AKF and MHSR together yields a 5.9-point OK-VQA decrease, which is 0.8 points larger than the sum of their separate decreases. This observation is descriptive rather than causal, but it is consistent with the design premise that repeated reasoning benefits from a cleaner evidence set.

2. Related Work

Knowledge-based VQA spans four threads relevant to KEAF-Net: attention-based grounding, knowledge-enhanced retrieval, graph-based reasoning, and large vision-language modeling. The central distinction is not the use of external knowledge alone, but how evidence quality is controlled before and during reasoning. Table 1, therefore, compares representative methods along five design axes: filter strategy, fusion mechanism, reasoning depth, filter supervision, and visual grounding. The comparison makes the research gap explicit: none of the listed baselines combines learned leave-one-out evidence supervision, typed heterogeneous fusion, and a three-hop recurrent reasoning stage within the same pipeline.

Table 1. Novelty comparison with recent KB-VQA methods

Method	Filter	Fusion	Reasoning	Filter Supv.	Grounding
MGRN [43]	Top-K graph	Contextual KG	Multi-hop	Task loss	Yes
RK-VQA [13]	Rational	FiD cross-attn	Decoder	Task loss	Limited
CEIK [12]	Gated	Softmax gate	Dual	Task loss	Limited
KEAF-Net	Learned	Het. graph	3-hop	LOO loss	Yes

2.1. Attention-based Visual Question Answering

Question-conditioned attention is widely used to link linguistic cues with relevant image regions. Anderson et al. [21] established the value of object-proposal features with top-down conditioning; MCAN [22] organized repeated co-attention between words and regions; and BAN [23] modeled fine-grained region-word interactions through bilinear attention. More recent VQA systems extend these ideas with graph attention, enhanced self-attention, cross-modal matching, and objectives intended to reduce language bias [24-28]. Surveys of robust VQA nevertheless emphasize a basic limitation: attention can reweight visible evidence, but it cannot supply a fact that is absent from the scene [29]. KEAF-Net, therefore, combines visual attention with an explicit external-knowledge path.

2.2. Knowledge-Enhanced Visual Question Answering

Knowledge-based VQA initially connected detected visual concepts to structured facts, as in FVQA [30] and later expanded to larger entity-centered settings such as KVQA. [31] Subsequent systems combined symbolic and neural reasoning, checked candidate answers against several sources, or injected retrieved knowledge into transformer decoders. [18, 32-34] Recent work has paid greater attention to question intent, evidence noise, and structured commonsense integration [11-13, 20]. A parallel line uses LLMs to provide background knowledge or reasoning traces, including PICa, Prophet, and PromptCap; [14-16] retrieval-augmented and late-interaction variants have also been studied [35]. The distinguishing choice in KEAF-Net is the training signal used for the triplet filter: relevance is estimated from the change in prediction loss when a triplet is removed rather than from answer-string overlap.

2.3. Graph-based Multimodal Reasoning

Graphs can represent objects, words, and facts as linked entities rather than as a single flattened feature vector. Surveys report the steady adoption of graph neural networks in vision-language tasks [36, 37]. Mucko used modality-specific graph convolutions for fact-based VQA; [38] VQA-GNN linked scene graphs with ConceptNet; [39] and heterogeneous or relation-aware graph variants have been explored in multimodal reasoning. [40-42] Multi-hop graph reasoning appears in MGRN, [43] NeSyGRN, [44] and MSG-KRM [45]. KEAF-Net places the filter before graph construction and uses a single typed graph for visual, textual, and knowledge nodes.

2.4. Large Vision-Language Models

Large pre-trained multimodal models, including BLIP-2 [3], LLaVA [4], InstructBLIP [5], and Flamingo [46], demonstrate the benefit of scaling visual-language pre-training and instruction following.

Knowledge-intensive VQA has also used question-aware captions, hierarchical retrieval, and knowledge filtering around large backbones [47-49]. These approaches provide strong background knowledge, but their billion-scale inference cost and opaque evidence use can make explicit support difficult to inspect.

KEAF-Net targets a complementary setting: a smaller inference model that retains selected triplets, preserves typed graph structure, and exposes hop-wise attention during reasoning.

3. Materials and Methods

For each image-question pair, KEAF-Net builds visual and textual representations, retrieves candidate knowledge, filters the retrieved triplets, fuses the retained evidence in a heterogeneous graph, and performs multi-hop answer reasoning. The trainable components are the filter, graph fusion layer, reasoning module, and classifier. Detector outputs and triplet sentence embeddings are pre-extracted so that training focuses on evidence selection and reasoning.

3.1. Architecture Overview

For an input image I and question Q, KEAF-Net predicts the answer from the visual representation V, textual representation T, and filtered triplet set K_hat, as formalized in Equation (1). Figure 1 summarizes the processing stages.

$$\hat{A} = Classifier(MHSR(HGAF(V, T, \hat{K})) \quad (1)$$

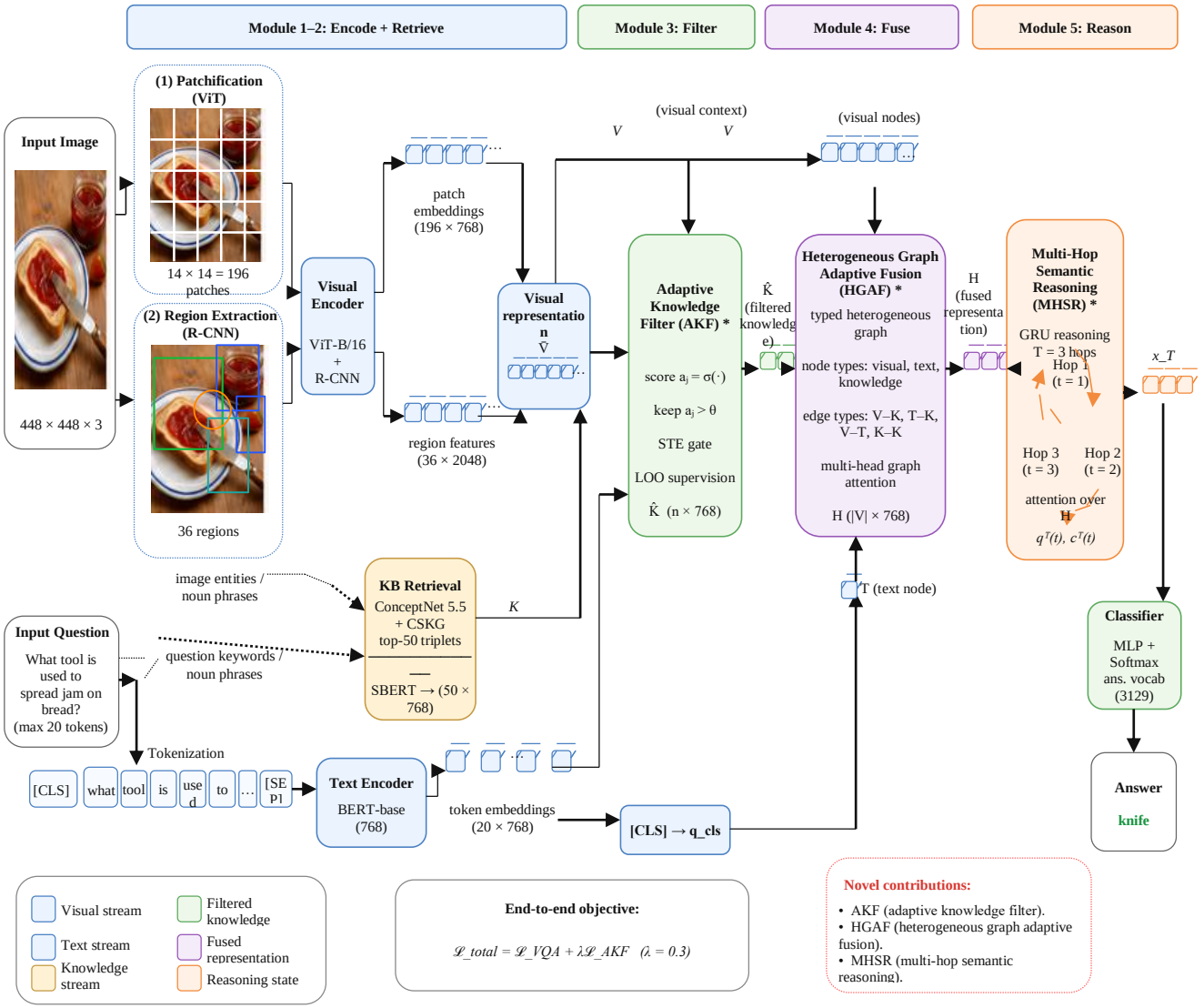

Fig. 1 KEAF-Net architecture with AKF, HGAF, and MHSR

3.2. Dual-Stream Feature Extraction

3.2.1. Visual Encoder

Images are partitioned into N_p patches by a ViT-B/16 [50] pre-trained on ImageNet-21K (86M parameters; layers 1-6 frozen), producing the grid features in Equation (2). A Faster R-CNN detector trained on Visual Genome [51] produces M = 36 region proposals (Equation (3)), which are concatenated with the grid features in Equation (4).

$$V_grid = ViT(I) = \{v_1, \dots, v_Np\} \in \mathbb{R}\^\{Np \times d\} \quad (2)$$

$$V_region = \{(r_1, b_1), \dots, (r_M, b_M)\} \quad (3)$$

$$V = [V_grid ; V_region] \in \mathbb{R}\^\{(Np + M) \times d\} \quad (4)$$

3.2.2. Textual Encoder

BERT-base [52] (110M parameters; embedding layer frozen) tokenizes and encodes the question, as shown in Equation (5).

$$T = BERT(Q) = \{t_1, \dots, t_L\}, q_cls = t_[CLS] \quad (5)$$

3.3. External Knowledge Retrieval

Retrieval starts from the top-10 Faster R-CNN labels above a confidence threshold of 0.3 and from question noun phrases extracted through spaCy chunking. These entities query a merged ConceptNet 5.5 [9] and CSKG [10] index. One-hop and two-hop neighbors are ranked by edge weight, duplicate entries are removed, and the list is capped at P = 50 triplets (Equation (6)). Each triplet is linearized and embedded with Sentence-BERT (all-MiniLM-L6-v2; pre-extracted offline), [53] as shown in Equation (7).

$$K = \{(s_j, r_j, o_j)\}, j = 1, \dots, P \quad (6)$$

$$k_j = SentBERT(s_j, r_j, o_j) \in \mathbb{R}\^d \quad (7)$$

3.4. Adaptive Knowledge Filter (AKF)

The retrieval stage favors recall, so the candidate set contains both helpful evidence and distracting facts. Because irrelevant facts can reduce accuracy, [13, 17] AKF is placed before graph construction. A naive answer-overlap label creates a shortcut that is unavailable at inference; therefore, KEAF-Net avoids dependence on the answer string. The image-question context is compressed into a single vector (Equation (8)), a three-way interaction scores each triplet (Equation (9)), and triplets above the learned threshold c_th are retained (Equation (10)).

$$h_IQ\ LayerNorm(W_q\ q_cls + W_v\ MeanPool(V)) \quad (8)$$

$$u_j = sigmoid(W_a[k_j ; h_IQ ; Had(k_j, h_IQ)] + b_a) \quad (9)$$

$$K_hat = \{k_j \mid u_j > c_th\} \quad (10)$$

Supervision uses a local counterfactual test: removing a helpful triplet should increase the VQA loss (Equation (11)). The loss difference is converted into a soft target through a temperature-scaled sigmoid (Equation (12)), and the filter loss matches predicted relevance to these targets (Equation (13)).

To bound training cost, |S| = 10 retained triplets are sampled per batch element for the leave-one-out target computation. Figure 2 details the AKF scoring path.

$$d_j = L_VQA(K_hat \setminus \{k_j\}) - L_VQA(K_hat) \quad (11)$$

$$t_j = sg(sigmoid(d_j / T_temp)), T_temp = 0.1 \quad (12)$$

$$L_AKF = mean_\{j \in S\}\,(u_j - t_j)^2 \quad (13)$$

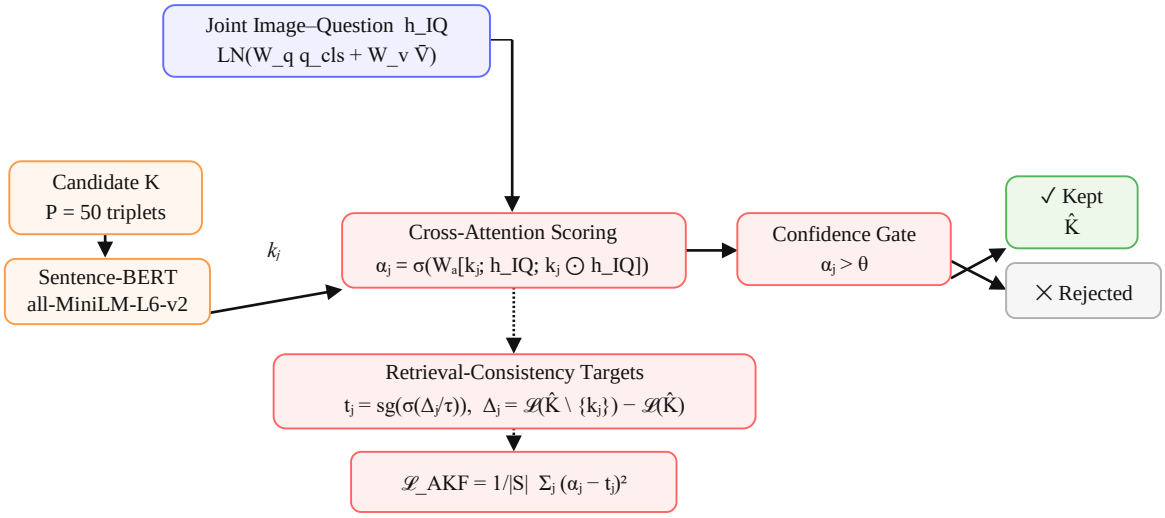

Fig. 2 Adaptive Knowledge Filter (AKF) with LOO supervision

3.5. Heterogeneous Graph Adaptive Fusion (HGAF)

HGAF constructs a graph G = (V, E) from three node families: visual regions, textual tokens, and retained knowledge triplets. Five connectivity rules determine the edges.

Visual-visual links use region IoU > 0.3; textual-textual links connect tokens within a window of 3; visual-knowledge links match detector labels to triplet entities; textual-knowledge links apply the same matching principle to noun phrases; and visual-textual links use cosine similarity > 0.4 after projection to the shared d-dimensional space.

A 500-pair manual link audit reported 82.4% precision and 71.6% recall for visual-knowledge links and 85.1% precision and 74.3% recall for textual-knowledge links. These values describe the heuristic linker and are separate from the AKF supervision diagnostics. Algorithm 1 summarizes graph construction and message passing, and Figure 3 illustrates the module.

Algorithm 1: HGAF Graph Construction and Fusion

Require: V, T, filtered knowledge K_hat

1: Init nodes V_v:=V, V_t:=T, V_k:=K_hat
2: Build edges E_vv,E_tt,E_vk,E_tk,E_vt (5 rules)
3: for l=1..L: node-type-conditioned attention over E (Equations (14) and (15))
4: h'_i = sigmoid(Σ_j p_ij W_type(j) h_j) (Equation (16))
5: gates g_v,g_t,g_k (Equation (17)); h_fused (Equation (18))
6: return h_fused

A multi-head GAT [54] with node-type-specific projections is applied over the adjacency induced by the five edge-construction rules (Equation (14)), followed by normalized attention (Equation (15)) and a sigmoid update (Equation (16)). The five edge categories determine which node pairs are connected; they are not presented as five independent relation-specific attention parameter sets.

Instead, the learned transformations are conditioned on node type, while the graph topology carries the relation-category information. Eight heads and two layers are used, and gated fusion combines the visual, textual, and knowledge streams (Equations (17) and (18)). This distinction keeps the mathematical specification consistent with the typed-neighborhood aggregation formulated in Algorithm 1.

$$e_ij = LeakyReLU(a^{T}[W_type(i)h_i \parallel W_type(j)h_j]) \quad (14)$$

$$p_ij = exp(e_ij) / \Sigma_k\, exp(e_ik) \quad (15)$$

$$h'_i = sigmoid(\Sigma_j\, p_ij\, W_type(j)\, h_j) \quad (16)$$

$$g_m = sigmoid(W_gm\, h_m + b_gm), m \in \{v, t, k\} \quad (17)$$

$$h_fused = \Sigma_m\, Had(g_m, h_m) \quad (18)$$

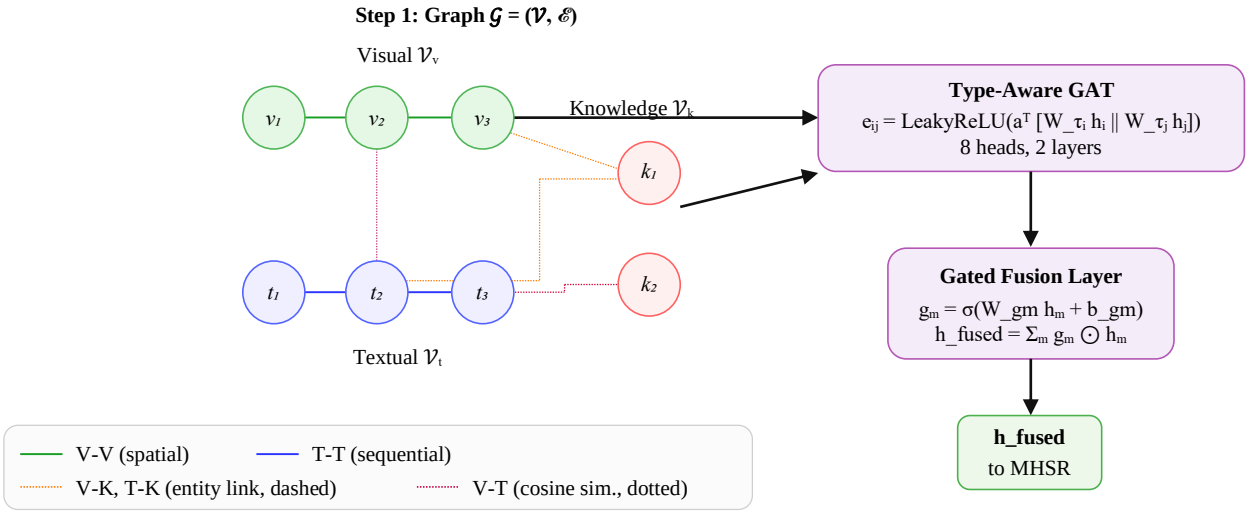

Fig. 3 Heterogeneous Graph Adaptive Fusion (HGAF)

3.6. Multi-Hop Semantic Reasoning (MHSR)

Knowledge-grounded questions often require chaining two or three pieces of evidence, making a single pass insufficient. MHSR iterates T hops over the fused graph (T = 3 by default), initializes the query from the [CLS] state (Equation (19)), attends over graph nodes (Equation (20)), and updates the query with a GRU (Equation (21)). The final query and attended evidence are concatenated for prediction (Equation (22)). Figure 4 shows the module.

$$q\hat{}(0) = q_cls \quad (19)$$

$$c\hat{}(t) = \Sigma_i\, \beta_i\hat{}(t) h_i,\ \beta_i\hat{}(t) = softmax(q\hat{}(t)^{T} W_hop\ h_i) \quad (20)$$

$$q\hat{}(t+1) = GRU(q\hat{}(t), c\hat{}(t)), t = 0, \ldots, T-1 \quad (21)$$

$$A_hat = softmax(W_out[q\hat{}(T)\,;\, c\hat{}(T)] + b_out) \quad (22)$$

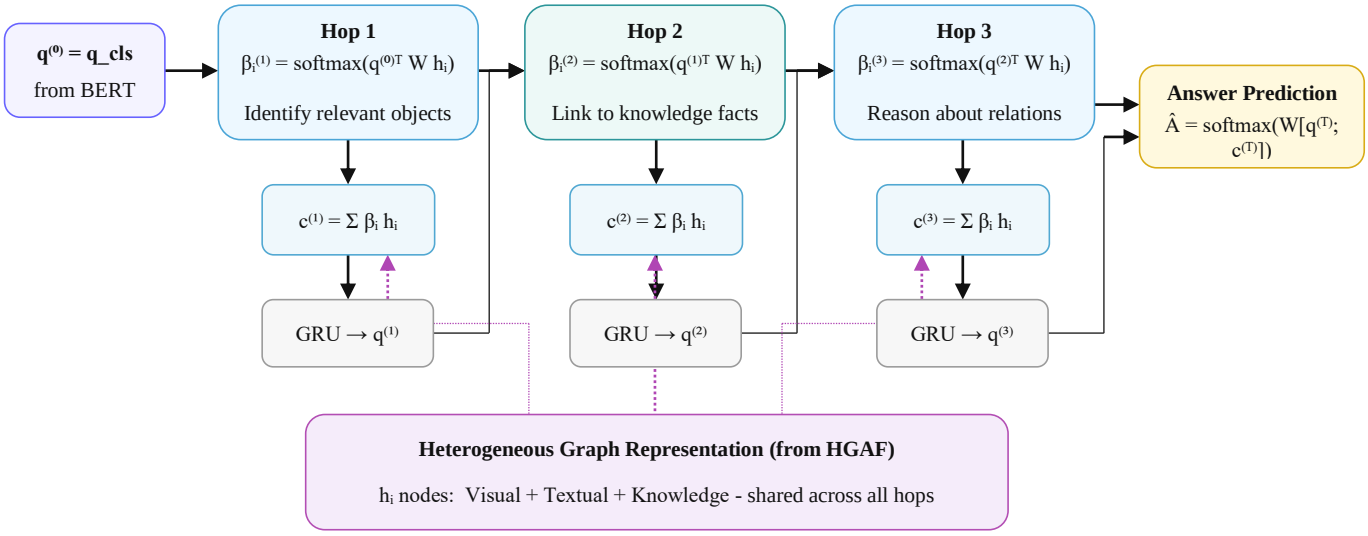

Fig. 4 Three-Hop Multi-Hop Semantic Reasoning (MHSR)

3.7. Training Objective

The total loss combines the VQA loss with the auxiliary filter loss (Equation (23)), where L_VQA is soft-target binary cross-entropy over the answer vocabulary and w_AKF = 0.3. Training is conducted for 20 epochs with AdamW and early stopping.

$$L_total = L_VQA + w_AKF\, L_AKF, w_AKF = 0.3 \quad (23)$$

4. Theoretical Analysis

The following analysis motivates AKF and HGAF, and the related claims are evaluated empirically through the ablation study.

4.1. AKF as an Information Bottleneck

Let X denote the raw triplets, Y the answer distribution, and Z the retained representation. The AKF design can be interpreted through the Information Bottleneck objective (Equation (24)): the threshold compresses the retrieval set, while the leave-one-out signal favors triplets whose presence improves prediction. [55, 56] This connection is interpretive rather than a claim that AKF is a variational optimizer.

$$min_Z\ I(X;Z) - beta\, I(Z;Y) \quad (24)$$

4.2. Leave-One-Out Marginal Contribution as a Data-Valuation Signal

The leave-one-out score in Equation (11) measures a local marginal contribution relative to the currently retained evidence set. This is related to the broader data-valuation perspective underlying Data Shapley, [57], but it is not equivalent to a Shapley value: Data Shapley averages marginal contributions over many coalitions, whereas the present signal evaluates the effect of removing one triplet from the current retained set. Full coalition enumeration is infeasible for P = 50 triplets, so sampled leave-one-out evaluation provides a tractable, mini-batch-compatible supervision signal. [58] The theoretical claim is therefore limited to local marginal valuation rather than full Shapley attribution.

4.3. Typed-Graph Representation in HGAF

Proposition 1 (Typed-neighborhood discrimination). A GAT with node-type-specific projections can distinguish neighborhood patterns that a type-agnostic GAT may merge when the neighborhoods contain different mixtures of visual, textual, and knowledge nodes. In HGAF, the five edge categories determine graph connectivity, while the node-type projections in Equation (14) condition message transformations on the participating modalities. This can be interpreted as refining neighborhood aggregation with type information; it does not claim general expressiveness beyond the 1-WL family [59, 60].

4.4. Computational Complexity and Depth

With hidden dimensionality fixed, AKF scoring grows linearly with the number of candidate triplets, HGAF message passing grows with the number of graph edges and attention layers, and MHSR grows with the number of reasoning hops. The sampled leave-one-out procedure affects training only and is not required at inference. The depth experiment in Table

4 provides the empirical control for recurrent reasoning: performance improves from one to three hops and then declines slightly at four and five hops. Accordingly, T = 3 is used as the default reasoning depth.

5. Experimental Setup

Experiments use four reported VQA benchmarks: VQA v2.0 [2] for general VQA, OK-VQA [6] and A-OKVQA [7] for knowledge-intensive VQA, and GQA [61] for compositional reasoning. OK-VQA has a public training split of about 9K questions and a 5,046-question evaluation split on COCO val2014; results follow the common protocol of reporting on this split. [15, 16, 33, 34, 46] A-OKVQA uses the official direct-answer test split. The primary metric is standard VQA soft accuracy (Equation (25)). KEAF-Net is trained three times with seeds {42, 123, 2024}, and tables report mean ± standard deviation. The classifier vocabulary contains the 3,129 most frequent answers.

$$Acc(\hat{A}) = min(1, (1/3)|\{a \in GT : a = \hat{A}\}|) \qquad (25)$$

5.1. Implementation and Reproducibility Configuration

The implementation uses ViT-B/16 with layers 1-6 frozen; Faster R-CNN region features with M = 36; BERT-base with the embedding layer frozen; ConceptNet 5.5 plus CSKG retrieval; Sentence-BERT all-MiniLM-L6-v2 triplet embeddings pre-extracted offline; P = 50 candidate triplets; 10 sampled leave-one-out evaluations per batch element with temperature 0.1; two HGAF layers with eight attention heads; three MHSR hops; AdamW optimization for 20 epochs with early stopping and effective batch size 64; a 3,129-answer vocabulary; and three runs with seeds {42, 123, 2024}. The reported 252M inference-parameter count refers to the answer model and excludes the offline Faster R-CNN and Sentence-BERT feature extractors. Exact hardware, learning-rate schedule, and wall-clock training time are not used for cross-paper efficiency claims because those metadata are not available in a common controlled protocol. All mathematical expressions are retained as editable Microsoft Word equation objects rather than rasterized equation images.

5.2. Statistical Reporting

All KEAF-Net configurations are reported over three independent runs as mean ± standard deviation. The three-run summaries quantify run-to-run variability, but no inferential significance claim is made from n = 3 alone. In addition, cross-paper baseline comparisons are treated as descriptive because the cited studies use different backbones, external knowledge sources, prompting regimes, and evaluation settings. The controlled evidence for component contribution is therefore taken from the within-model ablations in Table 3 and the reasoning-depth study in Table 4.

6. Results and Discussion

6.1. Main Results

Table 2 compares KEAF-Net with representative attention-based, explicit-knowledge, retrieval-augmented, LLM-assisted, and large-VLM baselines using values reported by the corresponding publications. On OK-VQA, KEAF-Net reports 68.7%, which is 4.6 percentage points above RK-VQA (64.1%) and 2.2 points above GeReA (66.5%) in the cited settings. On the A-OKVQA direct-answer test split, KEAF-Net reports 62.4%; this is 2.8 points above PromptCap (59.6%) but 0.9 points below GeReA (63.3%). On VQA v2.0, KEAF-Net reports 72.86%. The measured OK-VQA gain is consistent with two design choices that are independently supported by the controlled ablations: filtering retrieved triplets before graph propagation, since removing AKF costs 3.2 points, and preserving typed cross-modal interactions, since removing HGAF costs 3.8 points. The MHSR ablation further shows a 1.9-point contribution from iterative reasoning. These observations explain why the complete pipeline is stronger than its internal variants; they are not used to claim that architecture alone causes the cross-paper differences because the compared systems are not compute- or protocol-matched.

Table 2. Published comparison values for OK-VQA and A-OKVQA. KEAF-Net values are mean ± std over 3 runs. *GPT-3 175B at inference; †zero-shot; ‡few-shot. Model scales/settings are reported only as context and are not compute-matched; NR = Not Reliably reported in a comparable form.

Category	Method	Reported scale / setting	OK-VQA	A-OKVQA (DA)	VQA v2
Attention-based	MCAN [22]	66M	—	—	70.63
Attention-based	BAN [23]	112M	—	—	70.04
Explicit KB	KRISP [18]	155M	38.9	—	—
Explicit KB	MuKEA [19]	NR	42.6	—	—
Explicit KB	MSG-KRM [45]	NR	43.6	—	—
LLM-assisted KB	KAT (ensemble) [34]	T5-large 770M + GPT-3 175B	54.4	—	—
Knowledge coordination	CEIK [12]	NR	60.5	—	—
LLM-assisted KB	REVIVE (ensemble) [33]	T5-large 770M + GPT-3 175B	58.0	—	—
Retrieval/KB	RK-VQA [13]	~3B	64.1	—	—

LLM-assisted	PICa* [14]	GPT-3 175B	48.0	—	—
LLM-assisted	PromptCap* [16]	GPT-3 175B	60.4	59.6	—
LLM-assisted	Prophet* [15]	GPT-3 175B	61.1	55.7	—
Retrieval-augmented	RA-VQA-v2 + FLMR [35]	BLIP-2/T5-XL setting	62.1	—	—
Large VLM	BLIP-2† [3]	12.1B	45.9	—	65.0
Large VLM	Flamingo-80B† [46]	80B	50.6	—	—
Large VLM	Flamingo‡ (32-shot)	80B	57.8	—	—
MLLM-assisted	GeReA [47]	InstructBLIP-based	66.5	63.3	—
Proposed	KEAF-Net	reported 252M	68.7±0.2	62.4±0.3	72.86±0.2

6.2. Ablation Study and Diagnostic Evidence

Table 3 isolates the contribution of each proposed module. Removing AKF reduces OK-VQA from 68.7% to 65.5% (-3.2 points), indicating that unscreened retrieved knowledge materially degrades the complete configuration.

Removing HGAF lowers accuracy to 64.9% (-3.8 points), the largest single-module decrease, which supports retaining visual, textual, and knowledge representations as typed graph nodes rather than collapsing them into a homogeneous interaction. Removing MHSR lowers accuracy to 66.8% (-1.9 points), showing that repeated query updates provide an additional but smaller gain after evidence selection and fusion.

The joint AKF+MHSR ablation reaches 62.8%, a 5.9-point drop; this is 0.8 points larger than the sum of their separate decreases and is treated as descriptive evidence of complementarity rather than proof of a causal interaction.

The source experiments additionally report a 1.3-point gain for leave-one-out filter supervision over answer-overlap labels, a 1.6-point gain from adding CSKG to ConceptNet when AKF is active, and a reduction from 68.7% to 65.6% when HGAF is replaced by a homogeneous GAT. Together, these comparisons locate the main performance benefit in evidence control and typed fusion rather than in reasoning depth alone.

Table 3. Ablation accuracy after removing individual KEAF-Net modules (mean ± standard deviation, three runs)

Configuration	OK-VQA	A-OKVQA	VQA v2	GQA
Full KEAF-Net	68.7±0.2	62.4±0.3	72.86±0.2	65.8±0.3
w/o AKF	65.5±0.4	59.8±0.3	71.94±0.3	64.6±0.4
w/o HGAF	64.9±0.3	58.6±0.4	71.50±0.3	63.5±0.3
w/o MHSR	66.8±0.3	60.7±0.3	72.34±0.2	64.1±0.3
w/o Knowledge	63.2±0.4	56.4±0.5	71.80±0.3	63.9±0.3
w/o AKF+MHSR	62.8±0.5	55.9±0.4	71.22±0.3	62.7±0.4
Baseline	58.4±0.3	52.1±0.4	70.63±0.2	61.4±0.3

Table 4 examines MHSR depth independently of the module-removal study. Increasing the number of hops from one to three improves the reported accuracy on all four datasets, whereas four and five hops produce small declines. The selected operating point is therefore T = 3. In these experiments, a few query-state updates help combine evidence, but additional passes over the same fused graph do not provide further gains.

Table 4. Accuracy obtained with one to five MHSR reasoning hops

T	OK-VQA	A-OKVQA	VQA v2	GQA
1	66.8±0.3	60.7±0.3	72.34±0.2	64.1±0.3
2	67.9±0.2	61.6±0.3	72.62±0.2	65.1±0.3
3	68.7±0.2	62.4±0.3	72.86±0.2	65.8±0.3
4	68.5±0.3	62.1±0.3	72.78±0.2	65.6±0.3
5	68.2±0.3	61.8±0.4	72.65±0.3	65.3±0.3

To make the performance evidence directly editable and verifiable in Microsoft Word, Figure 5 has been reconstructed from numerical results already reported in Tables 3 and 4.

Panel (a) visualizes the OK-VQA ablation accuracy for the complete model and module-removal configurations, while panel (b) visualizes OK-VQA and A-OKVQA accuracy as the number of MHSR reasoning hops changes from one to five. The selected operating point T = 3 is the maximum reported value in both series of the depth study.

The AKF loss weight remains fixed at w_AKF = 0.3 as specified in the training objective; no additional measurement is introduced by the reconstructed charts.

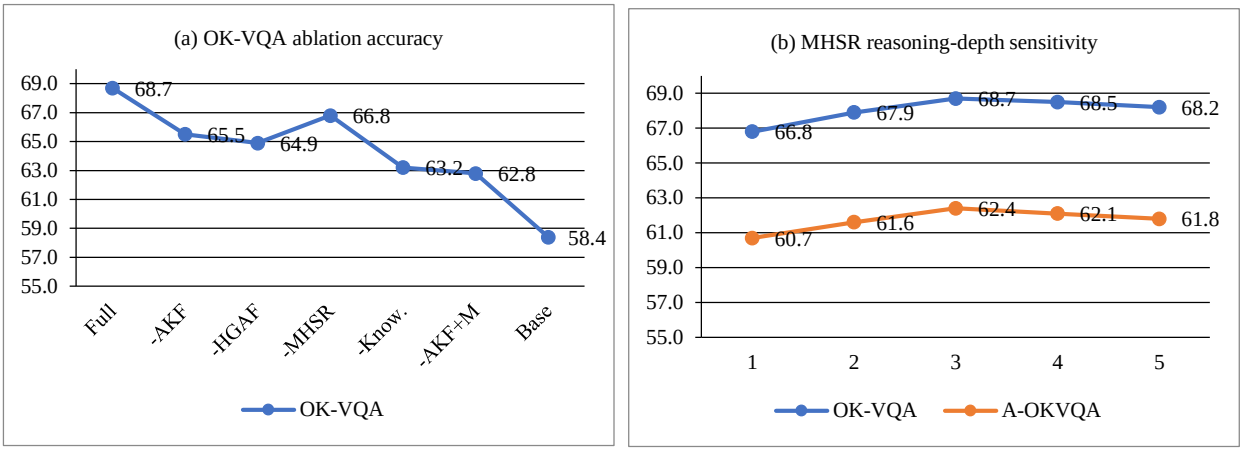

Fig. 5 Editable performance analysis: (a) OK-VQA ablation accuracy, and (b) Reasoning-depth sensitivity on OK-VQA and A-OKVQA.

6.3. Mechanistic Interpretation of the Ablations

The ablations can be interpreted in the same order in which the modules operate. AKF changes which external facts are admitted to the graph; HGAF controls how the retained visual, textual, and knowledge nodes exchange information; and MHSR determines how often the query is updated over that fused representation. The 3.8-point loss after removing HGAF is the largest single-module degradation, while the 3.2-point AKF loss shows that evidence quality must be controlled before graph propagation. Removing AKF and MHSR together lowers OK-VQA by 5.9 points, which is 0.8 points larger than the sum of their individual losses. This difference does not establish a causal interaction, but it is compatible with the intended dependency: iterative updates are most useful after distracting evidence has been reduced. The explanation is therefore grounded in within-model controls rather than in unmatched cross-paper comparisons.

6.4. Qualitative Analysis and Error Categories

Two validation examples illustrate the behavior expected from the filter-reasoner pipeline. For a question about the country of origin of green curry, the filtered configuration keeps a Thailand-related fact and removes a color-related distractor before predicting Thailand. For a koala question about its native continent, retaining an Australia-related fact changes the answer relative to the unfiltered configuration. These examples are qualitative checks rather than additional accuracy evidence. They also expose a practical limitation: when the retrieval source lacks the required temporal or entity-specific fact, downstream reasoning has no reliable evidence from which to recover the answer.

An exploratory review of 200 OK-VQA errors assigned 38% to retrieval failure, 24% to filter false negatives, 21% to insufficient reasoning depth, and 17% to visual or annotation errors. The first two categories together account for 62% of the inspected failures, placing most observed errors before the final multi-hop stage. Because the review was not repeated by a second annotator and no inter-annotator agreement statistic was collected, these percentages should be treated as diagnostic observations rather than a validated error taxonomy.

6.5. Limitations

Several constraints delimit the claims of this study. The model has no OCR-specific branch, so TextVQA-type questions [62] are outside the quantitative evaluation. Visual-knowledge and textual-knowledge links are produced by heuristic matching and can therefore be incorrect. AKF approximates evidence value with sampled leave-one-out effects rather than coalition-based valuation. The experiment archive available for this revision does not contain a complete hardware and optimizer-schedule record, which prevents exact runtime replication. Finally, comparisons with LLM-assisted and large-VLM systems use values reported under different training and inference settings; they provide context rather than controlled head-to-head evidence.

7. Conclusion

KEAF-Net combines adaptive triplet filtering, heterogeneous graph fusion, and multi-hop semantic reasoning for knowledge-intensive VQA. Across three runs, the model reports 68.7% on OK-VQA and 62.4% on A-OKVQA. Removing AKF, HGAF, and MHSR reduces OK-VQA by 3.2, 3.8, and 1.9 points, respectively, while the depth study peaks at three hops. These controlled results support the proposed ordering of operations: filter retrieved knowledge before message passing, preserve modality identity during fusion, and apply a short reasoning chain over the fused representation.

The cross-paper results are competitive within the cited settings but are not presented as a universal state-of-the-art claim across non-identical protocols. Future work should add learned entity linking, OCR-aware grounding, more complete reproducibility reporting, and richer evidence-valuation objectives while preserving a compact inference design.

References

[1] Stanislaw Antol et al., "VQA: Visual Question Answering," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425-2433, 2015. [Google Scholar] [Publisher Link]

[2] Yash Goyal et al., "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering," *International Journal of Computer Vision*, vol. 127, no. 4, pp. 398-414, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[3] Junnan Li et al., "BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models," *Proceedings of the 40th International Conference on Machine Learning*, pp. 19730-19742, 2023. [Google Scholar] [Publisher Link]

[4] Haotian Liu et al., "Visual Instruction Tuning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 34892-34916, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Wenliang Dai et al., "InstructBLIP: Towards General-Purpose Vision-Language Models with Instruction Tuning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 49250-49267, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Kenneth Marino et al., "OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 3190-3199, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[7] Dustin Schwenk et al., "A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge," *European Conference on Computer Vision*, Tel Aviv, Israel, pp. 146-162, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Jiaqi Deng et al., "A Comprehensive Survey of Knowledge-based Vision Question Answering Systems: The Lifecycle of Knowledge in Visual Reasoning Task," *arXiv preprint arXiv*, pp. 1-20, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[9] Robyn Speer, Joshua Chin, and Catherine Havasi, "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, pp. 4444-4451, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[10] Filip Ilievski, Pedro Szekely, and Bin Zhang, "CSKG: The CommonSense Knowledge Graph," *European Semantic Web Conference*, Springer, Cham, pp. 680-696, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Lingyun Song et al., "Answering Knowledge-based Visual Questions via the Exploration of Question Purpose," *Pattern Recognition*, vol. 133, pp. 1-12, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Qunbo Wang, Jing Liu, and Wenjun Wu, "Coordinating Explicit and Implicit Knowledge for Knowledge-based VQA," *Pattern Recognition*, vol. 151, pp. 1-9, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] Weipeng Chen et al., "RK-VQA: Rational Knowledge-Aware Fusion-in-Decoder for Knowledge-based Visual Question Answering," *Information Fusion*, vol. 118, pp. 1-12, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[14] Zhengyuan Yang et al., "An Empirical Study of GPT-3 for Few-Shot Knowledge-based VQA," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, pp. 3081-3089, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[15] Zhenwei Shao et al., "Prompting Large Language Models with Answer Heuristics for Knowledge-based Visual Question Answering," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, pp. 14974-14983, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Yushi Hu et al., "PromptCap: Prompt-Guided Image Captioning for VQA with GPT-3," *2023 IEEE/CVF International Conference on Computer Vision*, Paris, France, pp. 2951-2963, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Yanhan Sun et al., "DSAMR: Dual-Stream Attention Multi-Hop Reasoning for Knowledge-based Visual Question Answering," *Expert Systems with Applications*, vol. 245, pp. 1-9, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Kenneth Marino et al., "KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Knowledge-based VQA," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 14106-14116, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[19] Yang Ding et al., "MuKEA: Multimodal Knowledge Extraction and Accumulation for Knowledge-based Visual Question Answering," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, pp. 5079-5088, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[20] Shuo Yang et al., "MAGIC-VQA: Multimodal and Grounded Inference with Commonsense Knowledge for Visual Question Answering," *Findings of the Association for Computational Linguistics*, Vienna, Austria, pp. 16967-16986, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[21] Peter Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 6077-6086, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[22] Zhou Yu et al., "Deep Modular Co-Attention Networks for Visual Question Answering," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 6274-6283, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[23] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang, "Bilinear Attention Networks," *Advances in Neural Information Processing Systems*, vol. 31, pp. 1571-1581, 2018. [Google Scholar] [Publisher Link]

[24] Jing Lu et al., "Dual-Modality Integration Attention with Graph-based Feature Extraction for Visual Question and Answering," *Tsinghua Science and Technology*, vol. 30, no. 5, pp. 2133-2145, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[25] Rufai Yusuf Zakari et al., "Seeing and Reasoning: A Simple Deep Learning Approach to Visual Question Answering," *Big Data Mining and Analytics*, vol. 8, no. 2, pp. 458-478, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[26] Abdulganiyu Abdu Yusuf et al., "Graph-Enhanced Visual Representations and Question-Guided Dual Attention for Visual Question Answering," *Neurocomputing*, vol. 614, pp. 1-15, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[27] Jianjian Cao et al., "Bilateral Cross-Modality Graph Matching Attention for Feature Fusion in Visual Question Answering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 3, pp. 4160-4171, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[28] Souvik Chowdhury, and Badal Soni, "R-VQA: A Robust Visual Question Answering Model," *Knowledge-based Systems*, vol. 309, pp. 1-16, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[29] Jie Ma et al., "Robust Visual Question Answering: Datasets, Methods, and Future Challenges," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5575-5594, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[30] Peng Wang et al., "FVQA: Fact-based Visual Question Answering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 10, pp. 2413-2427, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[31] Sanket Shah et al., "KVQA: Knowledge-Aware Visual Question Answering," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 8876-8884, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[32] Jialin Wu et al., "Multi-Modal Answer Validation for Knowledge-based VQA," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, pp. 2712-2721, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[33] Yuanze Lin et al., "REVIVE: Regional Visual Representation Matters in Knowledge-based Visual Question Answering," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10560-10571, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[34] Liangke Gui et al., "KAT: A Knowledge Augmented Transformer for Vision-and-Language," *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, Seattle, United States, pp. 956-968, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[35] Weizhe Lin et al., "Fine-grained Late-interaction Multi-modal Retrieval for Retrieval Augmented Visual Question Answering," *Advances in Neural Information Processing Systems*, vol. 36, pp. 22820-22840, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[36] Abdulganiyu Abdu Yusuf et al., "Graph Neural Networks for Visual Question Answering: A Systematic Review," *Multimedia Tools and Applications*, vol. 83, no. 18, pp. 55471-55508, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[37] Henry Senior et al., "Graph Neural Networks in Vision-Language Image Understanding: A Survey," *The Visual Computer*, vol. 41, no. 1, pp. 491-516, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[38] Zihao Zhu et al., "Mucko: Multi-Layer Cross-Modal Knowledge Reasoning for Fact-based Visual Question Answering," *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, Yokohama Yokohama, Japan, pp. 1097-1103, 2021. [Google Scholar] [Publisher Link]

[39] Yanan Wang et al., "VQA-GNN: Reasoning with Multimodal Knowledge via Graph Neural Networks for Visual Question Answering," *2023 IEEE/CVF International Conference on Computer Vision*, Paris, France, pp. 21525-21535, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[40] Jing Zhang et al., "Cross-Modal Heterogeneous Graph Reasoning Network for Visual Question Answering," *Neural Computing and Applications*, vol. 37, no. 22, pp. 17701-17721, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[41] Yihan Zhao et al., "Heterogeneous Interactive Graph Network for Audio-Visual Question Answering," *Knowledge-based Systems*, vol. 300, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[42] Sai Zhang et al., "Relation-Aware Heterogeneous Graph Network for Learning Intermodal Semantics in Textbook Question Answering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 9, pp. 11872-11883, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[43] Zihan Hu et al., "A Multi-Hop Graph Reasoning Network for Knowledge-based VQA," *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 3, pp. 1-23, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[44] Muhammad Junaid Khan et al., "Enhancing Visual Question Answering with Common Sense Knowledge: A Data-Driven Neurosymbolic Graph Routing Approach," *International Journal of Data Science and Analytics*, vol. 20, no. 7, pp. 6391-6406, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[45] Lei Jiang, and Zuqiang Meng, "Knowledge-based Visual Question Answering using Multi-Modal Semantic Graph," *Electronics*, vol. 12, no. 6, pp. 1-19, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[46] Jean-Baptiste Alayrac et al., "Flamingo: A Visual Language Model for Few-Shot Learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23716-23736, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[47] Ziyu Ma et al., "GeReA: Question-Aware Prompt Captions for Knowledge-based Visual Question Answering," *arXiv preprint arXiv*, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[48] Davide Caffagni et al., "Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Seattle, WA, USA, pp. 1818-1826, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[49] Zhiyue Liu et al., "A Knowledge Noise Mitigation Framework for Knowledge-based Visual Question Answering," *2025 IEEE International Conference on Multimedia and Expo*, Nantes, France, pp. 1-6, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[50] Alexey Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *Proceedings of the International Conference on Learning Representations*, pp. 1-22, 2021. [Google Scholar]

[51] Ranjay Krishna et al., "Visual Genome: Connecting Language and Vision using Crowdsourced Dense Image Annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32-73, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[52] Jacob Devlin et al., "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Minneapolis, Minnesota, vol. 1, pp. 4171-4186, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[53] Nils Reimers, and Iryna Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Association for Computational Linguistics, Hong Kong, China, pp. 3982-3992, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[54] Petar Veličković et al., "Graph Attention Networks," *arXiv preprint arXiv*, pp. 1-12, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[55] Naftali Tishby, and Noga Zaslavsky, "Deep Learning and the Information Bottleneck Principle," *2015 IEEE Information Theory Workshop*, Jerusalem, Israel, pp. 1-5, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[56] Kenji Kawaguchi et al., "How Does Information Bottleneck Help Deep Learning?," *Proceedings of the 40th International Conference on Machine Learning*, pp. 16049-16096, 2023. [Google Scholar] [Publisher Link]

[57] Amirata Ghorbani, and James Zou, "Data Shapley: Equitable Valuation of Data for Machine Learning," *Proceedings of the 36th International Conference on Machine Learning*, pp. 2242-2251, 2019. [Google Scholar] [Publisher Link]

[58] Yongchan Kwon, and James Zou, "Beta Shapley: A Unified and Noise-Reduced Data Valuation Framework for Machine Learning," *arXiv preprint arXiv*, pp. 1-24, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[59] Keyulu Xu et al., "How Powerful are Graph Neural Networks?," *arXiv preprint arXiv*, pp. 1-17, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[60] Christopher Morris et al., "Weisfeiler and Leman Go Neural: Higher-Order Graph Neural Networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 4602-4609, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[61] Drew A. Hudson, and Christopher D. Manning, "GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 6693-6702, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[62] Amanpreet Singh et al., "Towards VQA Models that can Read," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 8309-8318, 2019. [CrossRef] [Google Scholar] [Publisher Link]