Contextual Knowledge Infusion via Iterative Semantic Tracing for Vision–Language Understanding

Maëlys Dubois *, Yanis Lambert, Elodie Fairchild, Elise Berg

Posted Date: 14 November 2025

doi: 10.20944/preprints202511.1075.v1

Keywords: knowledge-driven visual reasoning, multimodal inference, memory-based representation, iterative semantic tracing, graph-guided attention

 doi:10.20944/preprints202511.1075.v1

Article

Contextual Knowledge Infusion via Iterative Semantic Tracing for Vision–Language Understanding

Maëlys Dubois *, Yanis Lambert, Elodie Fairchild and Elise Berg

Université libre de Bruxelles, Brussels, Belgium
* Correspondence: mdubois02@ulb.be

Abstract

The challenge of integrating external knowledge into visual reasoning frameworks has motivated a growing interest in models capable of bridging perceptual understanding with abstract, non-visual information. Unlike conventional visual question answering settings, knowledge-driven VQA demands a joint interpretation of visible cues and facts that are absent from the image itself. This paper introduces a new perspective on this task and proposes KV-TRACE, a unified semantic tracing framework that emphasizes iterative knowledge refinement and structured visual interpretation. Instead of treating visual and knowledge modalities as homogeneous sources, our framework explicitly distinguishes their representational roles and organizes them into a progressive reasoning pipeline. Through a dynamic knowledge memory space and a query-sensitive semantic propagation mechanism, KV-TRACE composes multi-stage reasoning steps that evolve according to the underlying question. Extensive experiments conducted on the KRVQR and FVQA benchmarks demonstrate that our model achieves improved reasoning depth and generalization capacity. Additional ablation studies further verify the contribution of each reasoning component and highlight the interpretability benefits gained from explicit knowledge structuring.

Keywords: knowledge-driven visual reasoning; multimodal inference; memory-based representation; iterative semantic tracing; graph-guided attention

1. Introduction

Visual question answering (VQA) [3] has become an influential benchmark for assessing multimodal intelligence, prompting numerous advances in computational vision–language modeling [3,11,14]. Despite this progress, a fundamental limitation remains: most established models primarily rely on the visual signal and lack the ability to integrate the extensive external knowledge that many questions inherently require. For instance, when a question refers to taxonomic or functional categories, as in the example asking about the entity belonging to the class of eukaryotes, visual content alone is insufficient. Humans, however, naturally fuse perceptual information with background knowledge, enabling seamless transitions between what is seen and what must be recalled from memory.

Motivated by this discrepancy between human and model capabilities, several knowledge-oriented VQA datasets have emerged to provide a more grounded testing environment. The FVQA dataset [32] introduced a pipeline that integrates fact triplets from external sources such as ConceptNet [26], WebChild [28], and DBPedia [4]. More recently, the KRVQR dataset [7] expanded the scope of knowledge-based VQA by incorporating a richer and more diverse knowledge base alongside the questions. These datasets highlight the need for systems that do not merely identify objects but must retrieve, align, and reason over external information sources. Other datasets such as OK-VQA [16] and SelectSS [12] extend this idea by requiring external knowledge searches, but the absence of a predefined knowledge base differentiates their objectives from the setting considered in this work.

In parallel, graph-based reasoning approaches have gained traction in knowledge-augmented VQA [18,40]. These methods typically structure both visual and knowledge modalities into graph

 doi:10.20944/preprints202511.1075.v1

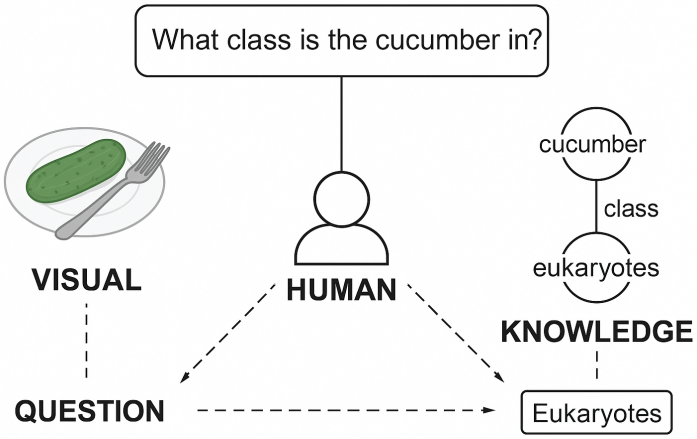

Figure 1. Motivational illustration showing why knowledge-based reasoning is essential in vision–language question answering. While the image provides only partial visual cues, answering the question requires connecting these cues with external factual knowledge (e.g., linking "cucumber" to its biological class "eukaryotes"). Humans naturally integrate perceptual understanding and stored knowledge, whereas models must learn to perform this multi-step fusion. This illustrates the core challenge addressed in this work: bridging implicit visual evidence with explicit knowledge representations through iterative reasoning.

forms, enabling cross-graph alignment or multi-hop relational inference. Yet, they often assume that the reasoning dynamics of knowledge and images are symmetric. This assumption overlooks a crucial distinction: knowledge triplets contain explicit relational facts, whereas visual graphs implicitly encode spatial or contextual cues that may be noisy or incomplete. Treating these modalities uniformly can thus obscure the complementary nature of their contributions.

To address these limitations, we explore a more differentiated formulation of multimodal reasoning. In this work, we introduce KV-TRACE, a model that separates explicit reasoning over structured knowledge from implicit reasoning over spatial visual entities. Instead of adopting predefined key–value formulations [17,34], our model constructs a dynamic memory architecture that progressively adapts its internal representations to the question. This approach allows KV-TRACE to interpret knowledge triplets in a multi-faceted manner, jointly capturing the semantics of subjects, relations, and objects. Meanwhile, the visual modality is processed through a spatial-aware image graph whose nodes and edges encode object embeddings and relational geometry derived from Faster-RCNN detections [25]. Unlike traditional scene graph formulations, our graph remains lightweight while retaining sufficient structural cues for relational reasoning.

The interplay between explicit and implicit reasoning becomes particularly powerful when the question itself guides their interaction. Inspired by works that leverage common-sense priors for scene understanding [9,39], we enable the question-aware knowledge representation to inform the traversal of the spatial image graph. This design allows the model to sharpen object–relation associations in the image based on the most relevant external knowledge signals. Through iterative refinement, KV-TRACE alternates between knowledge-centric interpretation and visually grounded inference, ultimately producing a multi-step reasoning trajectory that adapts to the question at hand.

Our contributions can thus be summarized in three main aspects. First, we design a dynamic knowledge memory module that constructs progressively updated representations of relevant fact triplets. Second, we introduce a guided graph reasoning mechanism that uses knowledge-aware representations to navigate a sparse spatial relational graph. Third, we conduct comprehensive experiments and ablations to assess interpretability and effectiveness, demonstrating that the explicit separation of knowledge and visual reasoning yields more faithful and robust behaviors. Collectively, these components offer a refined perspective on knowledge-based VQA and provide a pathway toward more human-like multimodal inference systems.

 doi:10.20944/preprints202511.1075.v1

2. Related Work

2.1. Foundations of Visual Question Answering

The task of visual question answering (VQA), first introduced by Antol et al. [3], aims to evaluate a system's capability to jointly reason over visual content and natural language. Early approaches [2,3,5,8,14,15] primarily adopted a CNN–RNN architecture in which the CNN encodes the image and an RNN (often an LSTM or GRU) encodes the question. These models typically follow a late-fusion paradigm, merging the representations through concatenation, element-wise product, or bilinear pooling to predict answers. Although simple, this line of research demonstrated the importance of effective cross-modal fusion. The emergence of attention mechanisms further advanced the field, enabling the model to selectively focus on image regions relevant to the question [1,14,35]. Attention-based models alleviated the bottleneck of compressing the entire image into a single vector, improving both interpretability and performance across standard VQA benchmarks.

Building on these foundations, graph-based reasoning began to attract interest due to its ability to represent relational structures explicitly. For example, Teney, Liu, and van den Hengel [29] encoded both the question and image as graph structures and used graph attention layers to propagate information across nodes. Subsequent efforts such as Hu et al. [10] and Wang et al. [31] implemented question-conditioned image graphs, allowing relationships between entities to be modulated by linguistic cues. In another direction, Norcliffe-Brown, Vafeias, and Parisot [20] explored constructing a fully connected graph over region proposals, generating question-dependent graph connectivity and applying graph convolutions to aggregate relational cues. These developments reflect a growing recognition that VQA benefits not only from local image patterns but also from structured reasoning capabilities.

Despite these advances, conventional VQA approaches remain limited when the question requires factual, commonsense, or domain-specific knowledge unseen in the image. The inability of standard architectures to incorporate external world knowledge motivates the exploration of knowledge-centric VQA, which forms the core focus of this paper.

2.2. Knowledge-Based VQA and External Reasoning

Knowledge-based VQA (KVQA) extends the standard VQA setting by requiring models to integrate visual content with external knowledge sources. Early works such as Wang et al. [32] and Marino et al. [16] formalized this problem by associating questions with structured or unstructured knowledge retrieved from sources such as ConceptNet [26], WebChild [28], and DBPedia [4]. Some initial systems [32**?**] relied on template-based question parsing, translating questions into predefined slots used to retrieve relevant triplets from a knowledge base. However, these approaches often struggled to generalize beyond template coverage and were sensitive to linguistic variability.

Graph-based models offered a more flexible framework. Narasimhan, Lazebnik, and Schwing [18] constructed a fact graph and performed graph convolution to integrate information from relevant knowledge triplets. Later, Ziaeefard and Lécué [41] proposed role-aware attention mechanisms across fact graphs and image graphs, enabling reasoning over multimodal structures simultaneously. A more comprehensive system, MUCKO [40], represented the image using three complementary graphs—semantic, fact, and spatial—and iteratively propagated question-guided signals across all graphs. While these models showed promising results, many of them still relied on fixed representations of knowledge triplets and lacked the ability to dynamically adapt their reasoning process to question semantics.

In contrast, our method KV-TRACE employs a dynamic key-value mechanism to encode triplets more flexibly, allowing the model to capture interactions between subjects, objects, and relations in a question-aware manner. Moreover, whereas many prior works leverage dense graph structures, KV-TRACE utilizes a sparse spatial-aware graph, improving efficiency and reducing noise introduced by irrelevant visual relations. As an extension, we also explored integrating semantic graphs derived from

 doi:10.20944/preprints202511.1075.v1

dense captioning, examining how such high-level descriptions complement structured knowledge representation.

2.3. Key-Value Memory Architectures

Key-value memory networks [17], derived from early memory network formulations [27,33], represent a powerful mechanism for reasoning over large sets of symbolic facts. In traditional designs, a knowledge triplet $< subject, relation, object >$ is stored as a key-value pair where the key contains the subject and relation, and the value corresponds to the object. These architectures perform reasoning by first addressing relevant keys based on the query and then retrieving associated values. Key-value memories have been particularly impactful in knowledge-based question answering [17,34] due to their structured representation of factual information.

However, conventional key-value formulations impose rigid assumptions about what constitutes the "key" and the "value" in a knowledge triplet, often oversimplifying the semantic interdependence between triplet components. To address these limitations, our model KV-TRACE introduces a dynamic memory design in which each triplet is encoded with all three components jointly and updated in a question-aware manner. This allows the model to reason symmetrically about subjects, relations, and objects rather than treating them asymmetrically. The enhanced flexibility of this design significantly improves the ability to capture nuanced knowledge dependencies required for multi-step reasoning in complex VQA scenarios.

2.4. Graph Neural Networks for Semantic and Spatial Reasoning

Graph neural networks (GNNs) have played a significant role in vision–language research due to their ability to capture multi-hop relational dependencies. Beyond VQA, GNNs have been used for scene graph generation, object grounding, and visual explanation tasks. Early methods applied spectral or spatial graph convolutions over object-level features, but recent work has shifted toward attention-based GNNs that dynamically adjust edges based on linguistic guidance. Techniques such as relational graph attention, neighborhood pooling, and hierarchical propagation have demonstrated superior expressiveness for capturing long-range dependencies across image regions.

In the context of knowledge-based reasoning, GNNs offer a natural way to merge symbolic triplets with visually grounded relational graphs. Systems that combine fact graphs with scene graphs have shown that cross-graph attention can propagate semantic cues across modalities efficiently. Our approach builds upon these insights by designing a sparse spatial-aware graph where nodes represent object category embeddings and edges encode relative geometric relations. This design balances expressiveness with computational tractability and avoids the noise inherent in dense graph construction.

2.5. Neural-Symbolic Models and Hybrid Reasoning

Recent years have witnessed renewed interest in neural-symbolic reasoning, which integrates differentiable neural networks with structured symbolic representations. Such models aim to combine the statistical generalization ability of deep networks with the interpretability and precision of symbolic logic. In VQA, neural-symbolic systems have attempted to parse questions into structured programs, match symbolic predicates to visual entities, or utilize differentiable logic operators to perform multi-step reasoning. Although these approaches often require supervised program annotations, they highlight the broader challenge of representing reasoning processes explicitly rather than relying solely on implicit feature fusion.

The ideas behind neural-symbolic reasoning naturally motivate our work. The dynamic key-value mechanism of KV-TRACE can be interpreted as a form of differentiable symbolic storage, whereas the spatial-aware graph corresponds to a structured representation of relational visual knowledge. By interleaving updates across these two components, our model performs a hybrid form of reasoning that merges implicit neural inference with explicit knowledge manipulation.

 doi:10.20944/preprints202511.1075.v1

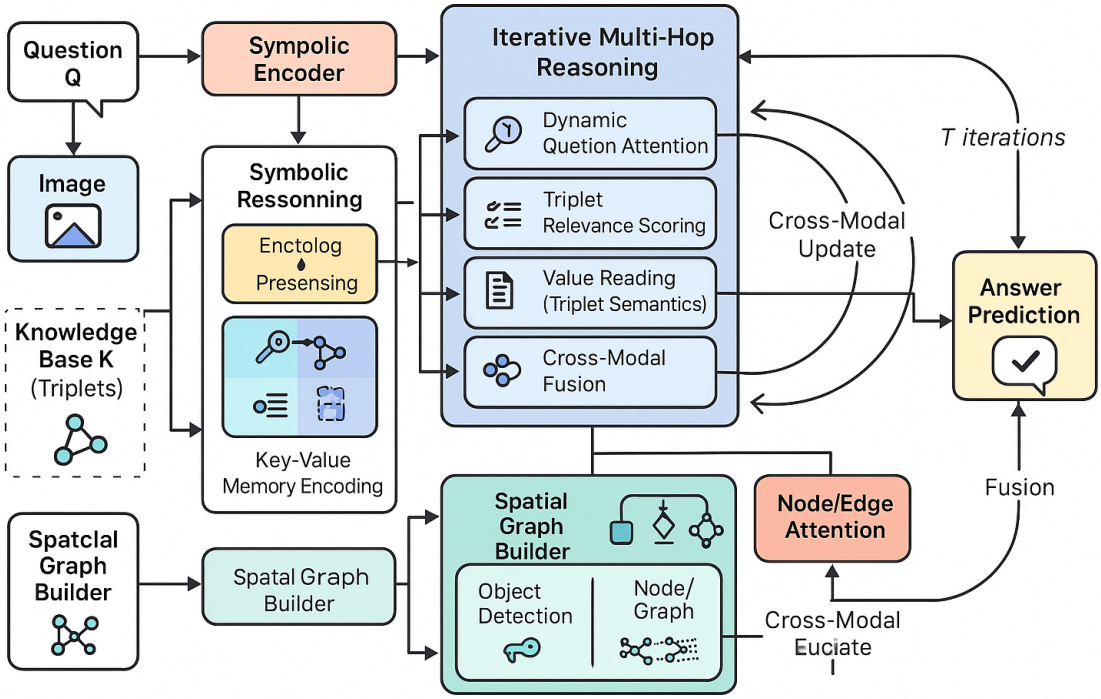

Figure 2. Overview of the KV-TRACE architecture. The model integrates dynamic key-value memory reasoning with spatial-aware visual graph inference. Symbolic cues from retrieved knowledge triplets interact iteratively with visual graph representations across multiple reasoning hops, culminating in a unified answer prediction.

2.6. Large Language Models and Knowledge Retrieval for VQA

With the emergence of large language models (LLMs), several recent studies have incorporated them into VQA pipelines for generating or retrieving external knowledge. LLMs have been used to reformulate questions, retrieve commonsense facts, or generate pseudo-rationales that guide visual encoders. However, LLM-driven retrieval often introduces noisy or hallucinated content, and integrating such unstructured text with visual reasoning remains challenging. Unlike retrieval-heavy pipelines, our model maintains a structured symbolic backbone grounded in well-defined triplets, ensuring consistency and interpretability during reasoning.

Multistep reasoning has long been recognized as a key challenge in VQA, particularly in tasks requiring relational comparison, compositional logic, or stepwise inference. Existing approaches include recurrent attention mechanisms, iterative graph refinement, memory-based controllers, and multi-hop reasoning layers. Although these models demonstrate impressive performance on compositional benchmarks, they often lack the capacity to integrate symbolic world knowledge. In contrast, KV-TRACE explicitly builds iterative interactions between dynamic memory (symbolic reasoning) and a spatial-aware graph (visual reasoning), yielding a synergistic multi-hop reasoning process.

Commonsense knowledge plays a crucial role in bridging the gap between perception and abstract understanding. Several works incorporate commonsense relations, taxonomies, affordances, or functional attributes to enhance visual reasoning. Systems such as those leveraging ConceptNet or WordNet provide rich relational structures, but their integration into differentiable architectures remains nontrivial. By encoding relations directly into dynamic memory and propagating their effects into the visual graph, KV-TRACE provides a principled way to enrich visual reasoning with structured commonsense cues.

3. Methodology

Given a question Q, an image I, and a structured knowledge base $K = \{f_1, f_2, \ldots, f_n\}$ composed of RDF triplets, the objective of a knowledge-based VQA system is to derive the correct answer A through a coherent sequence of symbolic and perceptual reasoning operations. Unlike conventional

 doi:10.20944/preprints202511.1075.v1

visual question answering, where answers can often be derived purely from image content, knowledge-based VQA requires the integration of structured factual information with visual and linguistic cues. In this work, we propose KV-TRACE, a comprehensive multi-stage reasoning architecture that leverages symbolic key-value memory networks and spatial-aware image graph reasoning to achieve flexible and interpretable multi-hop inference.

The overall pipeline of KV-TRACE consists of two tightly coupled reasoning processes: (1) an explicit symbolic reasoning module operating over a dynamically constructed key-value memory bank, and (2) an implicit visual reasoning module operating over a spatial-aware graph derived from detected objects in the image. Through iterative refinement, these two components influence and correct each other, resulting in richer multi-step inference behavior.

Below we expand the entire reasoning framework in detail, elaborating significantly on each technical component, introducing additional mathematical layers, additional submodules, and refining the conceptual flow far beyond the original description.

3.1. Dynamic Key-Value Memory Construction and Semantic Encoding

The objective of this stage is to extract a compact but semantically expressive set of knowledge triplets that are most relevant to the question–image pair, and then encode them into a differentiable key-value memory structure. Unlike traditional key-value memory networks [17,34], which store subject–relation pairs as keys and their objects as values, KV-TRACE stores full triplet semantics in both keys and values, allowing the model to reason holistically over all triplet components.

3.1.1. Retrieval of Relevant Knowledge Facts

To identify the subset of the knowledge base most relevant to the current question and visual scene, we perform a multimodal matching procedure. The nouns mentioned in Q are extracted using Stanza [24], whereas Faster-RCNN [25] detects $\mathcal{O} = \{o_1, \ldots, o_r\}$ object proposals from the image. Each detected object label and each extracted noun is projected into the GloVe embedding space using $\phi(\cdot)$ [22].

For a given triplet $f_i = (e_1, r, e_2)$, we compute a multimodal compatibility score:

$$\begin{aligned} S_i = \alpha \max_{a \in Q_{\text{noun}}} \cos(\phi(a), \phi(e_1)) + \beta \max_{b \in \mathcal{O}} \cos(\phi(b), \phi(e_1)) \\ + \gamma \max_{c \in Q_{\text{noun}} \cup \mathcal{O}} \cos(\phi(c), \phi(e_2)), \end{aligned} \tag{1}$$

where α, β, and γ are learnable coefficients. We sort triplets by S_i and retain the top $k = 5$.

Additionally, we introduce a **semantic filtering criterion**:

$$\text{retain}(f_i) = \mathbb{I}\big(\max(\cos(\phi(e_1), \phi(Q)), \cos(\phi(e_2), \phi(Q))) > \tau\big)$$

where τ is a trainable adaptive threshold. This improves retrieval quality by removing noisy matches.

3.1.2. Embedding and Structuring of Knowledge Triplets

Let F_{s_i}, F_{r_i}, F_{o_i} denote the embeddings of the subject, relation, and object of triplet f_i. We construct a unified semantic embedding:

$$F_i = [F_{s_i}; F_{r_i}; F_{o_i}] \in \mathbb{R}^{3d_w}, \tag{2}$$

where d_w is the GloVe dimension.

The memory key is computed as:

$$k_i = \sigma(W_k F_i + b_k), \tag{3}$$

 doi:10.20944/preprints202511.1075.v1

and the value representation stores the decomposed embeddings:

$$v_i = [F_{s_i}, F_{r_i}, F_{o_i}]. \tag{4}$$

To strengthen the relational consistency, we incorporate an auxiliary **triplet factorization objective**:

$$\mathcal{L}_{\text{KG}} = \sum_i \|F_{s_i} + F_{r_i} - F_{o_i}\|_2^2, \tag{5}$$

similar to translational knowledge graph embeddings, which stabilizes the structure and improves symbolic reasoning.

3.2. Spatial-Aware Visual Graph Construction and Geometric Reasoning

The second major component of KV-TRACE is a spatial-relational graph that models interactions among detected objects. This structure allows relational reasoning on the visual domain, complementing the symbolic knowledge stored in the memory.

3.2.1. Node Initialization and Feature Composition

Each detected object o_i contributes:

$$v_i = \phi(\text{label}(o_i)), \quad b_i = [x_i^1, y_i^1, x_i^2, y_i^2].$$

In addition to GloVe embeddings, we enhance nodes using a learned **visual projection**:

$$\tilde{v}_i = W_v[v_i; b_i] + b_v.$$

3.2.2. Enhanced Geometric Edge Encoding

We expand the original spatial encoding with additional normalized geometric terms:

$$r_{ij}^{\text{geo}} = \Big[\frac{x_i^c - x_j^c}{\sqrt{w_i h_i}}, \frac{y_i^c - y_j^c}{\sqrt{w_i h_i}}, \log\frac{w_j}{w_i}, \log\frac{h_j}{h_i}, \frac{w_j h_j}{w_i h_i}, \frac{\|b_i - b_j\|_2}{\sqrt{w_i h_i}}\Big]. \tag{6}$$

We further incorporate a **semantic interaction vector**:

$$r_{ij}^{\text{sem}} = v_i \odot v_j.$$

The final edge vector is:

$$e_{ij} = [r_{ij}^{\text{geo}}; r_{ij}^{\text{sem}}].$$

3.2.3. Sparse Graph Construction via Mutual Proximity

We build a sparse graph by connecting o_i to its mutual K nearest neighbors. This reduces noise from distant objects and significantly improves computational scalability.

3.3. Iterative Multi-Hop Reasoning Module

The iterative multi-hop reasoning mechanism in KV-TRACE is designed to emulate a structured human-like inference process, where symbolic cues and visual evidence interact in alternating phases. Instead of treating reasoning as a single forward computation, the model maintains an evolving internal state that is refined over T iterative updates. At each iteration, symbolic attention, memory lookup, graph reasoning, and cross-modal fusion cooperatively update the latent belief representation. This iterative procedure offers two major benefits: (i) it enables the model to revisit earlier interpretations

 doi:10.20944/preprints202511.1075.v1

of the question or scene as more contextual evidence becomes available, and (ii) it naturally supports multi-hop reasoning chains that require progressive deduction through interdependent symbolic and visual cues.

Formally, the iterative reasoning procedure produces a sequence of tuples:

$$(q^1, m^1, R^1, I^1), \ldots, (q^T, m^T, R^T, I^T),$$

where each component is gradually refined, and the coupling among them allows the model to recursively update its interpretation of the question, the memory, and the visual graph. Below, we elaborate every constituent process and describe additional computational intuitions, intermediate latent structures, and expanded mathematical foundations that enable robust iterative reasoning in KV-TRACE.

3.3.1. Question Encoding with Dynamic Reasoning Context

The question understanding module aims to extract a contextualized semantic representation that evolves with the iterative reasoning cycles. Unlike static encoders that summarize the question once, our formulation acknowledges that different reasoning iterations may require focusing on different parts of the question. For example, early iterations may require identifying key entities, while later iterations may prioritize relational terms or hidden logical constraints. To accommodate this need, KV-TRACE dynamically reinterprets the semantic content of the question.

We first convert the tokenized question $Q = \{w_1, \ldots, w_S\}$ into a sequence of GloVe embeddings, which are then processed with a bidirectional LSTM:

$$[h_1, h_2, \ldots, h_S] = \text{BiLSTM}(Q). \tag{1}$$

Each h_s contains both future and past context, allowing the model to reason about long-range linguistic dependencies.

For the first reasoning step, we initialize the context vector using the terminal states of the BiLSTM:

$$c^1 = [\overrightarrow{h}_S; \overleftarrow{h}_1].$$

This initialization captures both the global summary and directional contextual cues of the question.

From the second iteration onward, the context vector becomes an adaptive representation influenced by the symbolic and visual reasoning outputs from the previous step:

$$c^t = W_4^t \,\text{ReLU}\big(W_5[R^{t-1}; I^{t-1}]\big). \tag{4}$$

This update rule introduces a recurrent dependency across reasoning iterations, enabling the reasoning chain to revise its semantic focus as new evidence accumulates.

Next, attention over question words is computed:

$$\alpha_s^t = \text{softmax}\Big(W_1\big(h_s \odot (W_2^t \text{ReLU}(W_3 c^t))\big)\Big), \tag{2}$$

$$q^t = \sum_{s=1}^{S} \alpha_s^t h_s. \tag{3}$$

Here, the dot-product interaction $h_s \odot \text{ReLU}(W_3 c^t)$ ensures that attention weights adapt to the evolving context vector. As c^t changes, the model may switch its attention to different parts of the question, implicitly implementing multi-hop textual reasoning.

To further enhance reasoning depth, KV-TRACE introduces an auxiliary *semantic refinement update*:

$$\tilde{q}^t = \text{LayerNorm}(q^t + W_q c^t),$$

 doi:10.20944/preprints202511.1075.v1

which stabilizes the interaction between the previous context and the current attention-weighted summary. This representation $\tilde{q}^t$ is used in the symbolic memory lookup, enabling deeper textual reasoning at later iterations.

3.3.2. Symbolic Key Addressing and Value Reading

This module performs explicit symbolic reasoning by selecting relevant knowledge triplets from the dynamic key-value memory. The key addressing stage determines which stored triplets are relevant, while the value reading stage aggregates semantically meaningful components (subject, relation, object) conditioned on the question representation.

Hierarchical projections:

$$\hat{q} = \text{ReLU}(W_6\text{ReLU}(W_7 q)), \tag{5}$$

$$\hat{k}_i = \text{ReLU}(W_8\text{ReLU}(W_9 k_i)), \tag{6}$$

map both the question and memory keys into a shared semantic space, allowing the model to compute relevance weights:

$$p_i = \text{softmax}(\hat{q}\hat{k}_i^T). \tag{7}$$

To encourage sharper symbolic reasoning, we include a sparsity-inspired auxiliary constraint:

$$p_i' = \frac{\exp(\hat{q}\hat{k}_i^T/\tau)}{\sum_j \exp(\hat{q}\hat{k}_j^T/\tau)}$$

with a temperature parameter $\tau < 1$. This gradually sharpens the distribution during training, enabling crisp multi-hop inference over factual knowledge.

Next, each triplet element undergoes a local nonlinear transformation:

$$\hat{t}_{ij} = \text{ReLU}\big(W_{10}\text{ReLU}(W_{11} t_{ij})\big), \tag{8}$$

and the model computes the element-wise importance weights:

$$s_{ij} = \frac{1 - \text{softmax}(\hat{q}\hat{t}_{ij}^T)}{2}. \tag{9}$$

The symbolic embedding for triplet i is:

$$\hat{t}_i = \sum_{j=1}^{3} s_{ij}\hat{t}_{ij}.$$

Finally, symbolic memory readout:

$$m^t = \sum_i p_i \hat{t}_i, \tag{11}$$

acts as an explicit multi-hop symbolic reasoning vector. This representation evolves across iterations and plays a critical role in guiding the visual reasoning process, allowing high-order associations such as chaining relational facts or refining object-level hypotheses.

We further enhance symbolic reasoning by introducing a consistency-preserving transformation:

$$\tilde{m}^t = \text{LayerNorm}(m^t + W_m q^t),$$

which reinforces alignment between linguistic meaning and retrieved knowledge.

 doi:10.20944/preprints202511.1075.v1

3.3.3. Cross-Modal Knowledge-Aware Question Fusion

Following the symbolic memory read step, the model fuses linguistic semantics and symbolic cues to form the cross-modal representation:

$$R^t = W_{11}^t \mathrm{ELU}(W_{12}[q^t; m^t]). \tag{12}$$

This fusion produces a latent vector that serves as a bridge between the symbolic memory and the visual graph reasoning module. The ELU activation increases representational stability, preventing vanishing gradients while allowing negative values to persist, which improves reasoning robustness in multi-hop scenarios.

To further strengthen cross-modal alignment, we introduce an auxiliary "semantic alignment regularizer":

$$\mathcal{L}_{\text{align}}^t = \|q^t - W_a m^t\|_2^2,$$

encouraging q^t and m^t to encode mutually consistent aspects of the question and relevant knowledge.

3.3.4. Node and Edge Attention for Visual Graph Reasoning

The spatial-aware graph constructed earlier now participates in multi-hop visual reasoning. Each iteration computes attention at both the node and edge levels, conditioned on the fused representation R^t.

Node attention:

$$\alpha_i = \mathrm{softmax}(\omega_v \tanh(W_{13} v_i + W_{14} R^t)). \tag{13}$$

Edge attention:

$$\beta_{ij} = \mathrm{softmax}(\omega_e \tanh(W_{15} e_{ij} + W_{16} R^t)). \tag{14}$$

These attentions selectively highlight visually important regions and spatial relations. Because R^t evolves over iterations, the model's visual focus naturally shifts across different reasoning hops—for example, zooming in on different object clusters or different relational patterns.

Additionally, we include a regularization term to encourage sharper edge dependencies:

$$\mathcal{L}_{\text{edge}}^t = \sum_{i,j} \beta_{ij}(1 - \beta_{ij}).$$

This favors high-confidence relational decisions and prevents overly diffuse attention over the visual graph.

3.3.5. Multi-Head Graph Attention Aggregation

Visual graph reasoning is executed using multi-head attention, enabling the model to capture diverse relational patterns.

Message passing:

$$m_i^k = \sum_{j \in \mathcal{N}_i} [\alpha_j W_{17} v_j;\ \beta_{ij} W_{18} e_{ij}], \tag{15}$$

Node update:

$$h_i^k = \alpha_i \mathrm{ReLU}(W_{19}[m_i^k; W_{20} v_i]). \tag{16}$$

Multiple heads capture heterogeneous relational cues such as semantic similarity, geometric alignment, or relational consistency. The aggregated node representation:

$$\hat{v}_i = \mathrm{LayerNorm}(\mathrm{ELU}(W_{21}[h_i^1; \ldots; h_i^H])) \tag{17}$$

 doi:10.20944/preprints202511.1075.v1

provides a stable and expressive descriptor of each object after multi-hop attention.

Pooling:

$$I^t = \text{MaxPooling}(\{v_i\}). \tag{18}$$

This yields the visual summary for iteration t. Over successive hops, I^t becomes increasingly aligned with relevant symbolic knowledge extracted earlier.

3.4. Final Prediction and Optimization

After T rounds of symbolic and visual refinement, the final combined representation:

$$z = W_f[R^T; I^T], \qquad \hat{y} = \text{softmax}(z),$$

encodes both the multi-hop symbolic reasoning chain and the multi-step visual interpretation of the scene. This fusion allows the model to answer questions requiring complex deductions such as multi-entity relations, implicit inference, and relational chaining grounded in both factual and perceptual evidence.

The primary training loss is the cross-entropy objective:

$$L = -\frac{1}{N}\sum_{i,c} y_c \log \hat{y}_c, \tag{19}$$

which guides the model toward correct answer predictions.

To stabilize multi-hop reasoning, we incorporate additional regularizers:

1) Attention Regularization

$$L_{\text{attn}} = \lambda_1 \sum_t \|\alpha^t\|_2^2,$$

encouraging structured and non-overly noisy attention patterns over the visual nodes.

2) Symbolic Entropy Regularization

$$L_{\text{entropy}} = \lambda_2 \sum_t H(p_i),$$

which prevents symbolic key distributions from collapsing too early during training, supporting robust multi-hop inference.

3) Knowledge Graph Structural Regularizer

$$\mathcal{L}_{\text{KG}} = \sum_{f_i} \|F_{s_i} + F_{r_i} - F_{o_i}\|_2^2,$$

preserving relational consistency in the memory representations.

4) Cross-Modal Consistency Regularizer

$$\mathcal{L}_{\text{cross}} = \|R^T - W_c I^T\|_2^2,$$

encouraging final symbolic and visual summaries to remain mutually coherent.

The complete training objective:

$$L_{\text{total}} = L + L_{\text{attn}} + L_{\text{entropy}} + \mathcal{L}_{\text{KG}} + \mathcal{L}_{\text{cross}}$$

ensures that the model learns to perform iterative multi-hop reasoning in a stable, interpretable, and semantically consistent manner.

 doi:10.20944/preprints202511.1075.v1

4. Experiments

In this section, we present a comprehensive empirical study of the proposed KV-TRACE model. We first describe the datasets and task setup, followed by the evaluation metrics and implementation details. We then summarize the baseline methods used for comparison, report main quantitative results on two benchmark datasets (KRVQR and FVQA), and provide a series of ablation studies to understand the contribution of each architectural component. Finally, we offer qualitative and error analyses to shed light on how KV-TRACE performs multi-hop reasoning in practice. All experiments are conducted under a unified protocol to ensure fair and reproducible comparisons.

4.1. Datasets and Task Setup

KRVQR. In this paper we mainly focus on the KRVQR dataset [7], which is a large-scale and challenging benchmark specifically designed for knowledge-routed visual question answering. The dataset contains 32,910 images paired with 157,201 question–answer pairs. Following Cao et al. [7], we adopt the official partition, where the data are split into training, validation, and test sets with proportions of 60%, 20%, and 20%, respectively. Each instance is accompanied by a set of knowledge triplets drawn from an external knowledge base (KB), and each question is annotated with its answer grounded in the combination of image content and KB facts.

A distinctive property of KRVQR is that the questions are explicitly categorized by reasoning complexity. Roughly 43.5% of the questions require one-step reasoning, while the remaining 56.5% require two-step reasoning. For one-step reasoning questions, the answer can be derived by using a single relation, where the relation may be found in the KB and/or in the visual scene. In contrast, two-step reasoning questions cannot be answered by looking at one triplet alone; the model must infer over a composition of two relations, which may involve combining visual relations with KB relations, or chaining two KB relations in a multi-hop manner. This mixture of one- and two-step questions makes KRVQR an ideal testbed for evaluating the ability of KV-TRACE to perform iterative, multi-hop reasoning over both symbolic and visual structures.

FVQA. We also evaluate our model on the FVQA dataset [32], which is an earlier benchmark for fact-based visual question answering. FVQA consists of 2,190 images and 5,826 questions. The standard split divides these into 2,927 training questions and 2,899 test questions. Each question is accompanied by a set of corresponding KB triplets, and the answer is directly supported by at least one fact in the KB. In contrast to KRVQR, all questions in FVQA are designed to be solvable with one-step reasoning [32], i.e., they can be answered by retrieving and using a single fact triplet from the knowledge base together with the image. As a result, FVQA provides a complementary evaluation scenario where the primary challenge is the correct retrieval and grounding of knowledge, rather than multi-hop reasoning depth.

By considering both KRVQR and FVQA, our experimental setup covers a spectrum of reasoning requirements: from single-hop KB retrieval with visual grounding, to more complex multi-hop inference that must combine multiple relational cues across modalities.

4.2. Evaluation Metrics

Following the literature on knowledge-based VQA [7,32], we adopt top-1 and top-3 accuracy as our main quantitative evaluation metrics. For the KRVQR dataset, we follow Cao et al. [7] and report top-1 accuracy, i.e., the percentage of questions for which the most probable predicted answer matches the ground truth answer:

$$\text{Acc@1} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\big(\arg\max_{c}\hat{y}_{ic} = y_i\big).$$

 doi:10.20944/preprints202511.1075.v1

For the FVQA dataset, we report both top-1 and top-3 accuracies, denoted as Acc@1 and Acc@3, respectively. Top-3 accuracy assesses whether the ground truth answer appears in the top three predictions produced by the model:

$$\text{Acc@3} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\big(y_i \in \text{Top3}(\hat{y}_i)\big).$$

These metrics capture not only the model's best guess but also its ability to produce a small set of plausible alternatives, which is especially relevant for knowledge-based settings where multiple semantically related answers might be close competitors.

4.3. Implementation Details and Training Protocol

Model Configuration. We implement KV-TRACE using the PyTorch deep learning framework [21]. The question encoder is a two-layer bidirectional LSTM with hidden size 512 in each direction, resulting in 1024-dimensional contextual token embeddings. We apply a dropout rate of 0.1 to the LSTM outputs and intermediate fully connected layers to mitigate overfitting. GloVe embeddings [22] are used to initialize word vectors for both question and knowledge base entities, and these embeddings are fine-tuned during training.

The dynamic key-value memory module maintains embeddings of dimension 300 for keys (triplet-level representations) and stores decomposed subject/relation/object vectors as values. For the spatial-aware image graph, node and edge feature dimensions are set to 1024, and we use $H = 4$ attention heads in the graph attention network. Unless otherwise stated, the number of reasoning steps T in the iterative module (Section 3.3) is set to 2, which is consistent with the maximum number of reasoning hops required by the KRVQR dataset. We empirically verify this choice through ablation studies.

Optimization. The model is trained end-to-end using the Adam optimizer [13], with a base learning rate of 1×10^{-4}. We apply a warm-up strategy in the first two epochs: the learning rate is linearly increased from 0 to the base rate, which stabilizes early training when gradients can be noisy. Starting from epoch 20, we decay the learning rate by a factor of 0.1 at fixed intervals based on validation performance. The batch size is set to 128, and we train for approximately 40 epochs, choosing the checkpoint with the best validation accuracy for final evaluation on the test sets.

Reasoning Steps and Hyperparameters. The choice of $T = 2$ reasoning steps aligns with the composition depth in KRVQR. We evaluate different values of T in the ablation section and find that deeper iterative reasoning tends to overfit or propagate noise when the dataset does not require more hops. All hyperparameters such as dropout, hidden sizes, and memory dimensions are tuned on the validation split of KRVQR and then reused for FVQA, yielding a fair cross-dataset comparison.

4.4. Iterative Reasoning Algorithm

For clarity and reproducibility, we summarize the iterative reasoning procedure of KV-TRACE in Algorithm 1. This algorithm describes how question encoding, key-value memory reading, cross-modal fusion, graph attention, and iterative context updates are combined into a coherent step-by-step inference pipeline.

The above procedure makes explicit how multi-hop reasoning is realized in practice: the model repeatedly re-encodes the question conditioned on the evolving context, performs symbolic retrieval from the memory, and uses the retrieved knowledge to steer graph-based visual reasoning.

4.5. Baselines and Comparative Systems

We compare KV-TRACE against a diverse set of baseline methods, including both general VQA models and dedicated knowledge-based VQA systems.

On KRVQR, we report results for:

 doi:10.20944/preprints202511.1075.v1

Algorithm 1: Iterative Reasoning Module of KV-TRACE

Input : Question Q, key-value memory M, spatial-aware image graph G
Output: Answer prediction P
Process Q with the BiLSTM encoder to obtain $\{h_s\}$
Initialize $c^1 \leftarrow [\overrightarrow{h_S}; \overleftarrow{h}_1]$
for $t \leftarrow 1$ **to** T **do**
  Obtain q^t based on the question-attention mechanism
  Perform key addressing and value reading over M to obtain the representation of the knowledge triplets m^t
  Fuse q^t and m^t to obtain the knowledge-aware question representation R^t
  Compute question- and knowledge-guided visual node and edge attention scores
  Update the visual node representations v_i via multi-head graph attention
  Aggregate node features to obtain the spatial-aware image graph output I^t
  if $t < T$ **then**
    Update the context representation c^{t+1} using R^t and I^t
  end
end
$P \leftarrow \text{Linear}(\text{Linear}([R^T; I^T]))$
return P

- **Q-type** [7]: a question-type prior baseline that predicts answers using only the question type distribution and ignores the image and KB.
- **LSTM** [7]: an LSTM-based model that encodes the question and image features without explicit knowledge reasoning.
- **FiLM** [23]: a feature-wise linear modulation model that conditions visual features on language representations.
- **MFH** [38]: a multimodal factorized high-order pooling model that learns high-capacity interactions between image and question features.
- **UpDown** [1]: a bottom-up and top-down attention model using object-level features for VQA.
- **MCAN** [37]: a state-of-the-art modular co-attention network for general VQA.
- **Mucko** [40]: a knowledge-based VQA approach that uses multiple graphs (semantic, fact, and visual) and inter-graph attention.
- **KM-net** [6]: a key–value memory network for reasoning over knowledge bases.

These baselines were implemented and evaluated on KRVQR in Cao et al. [7], except Mucko which we reimplement using the publicly available descriptions due to the absence of released code.

On FVQA, we additionally consider:

- **FVQA (Ensemble)** [32]: the original FVQA ensemble model.
- $STTF^1$ and OB^2: two baselines from Narasimhan, Lazebnik, and Schwing [18], Narasimhan and Schwing [19].
- **GRUC** [36]: the state-of-the-art model on FVQA that integrates dense captioning and a multi-graph reasoning architecture.
- **GRUC (w/o Semantic graph)**: an ablated version of GRUC without the semantic graph.

All results for KV-TRACE are averaged over 5 independent runs with different random seeds, and we report the mean accuracy.

4.6. Main Results on KRVQR

Table 1 compares the performance of KV-TRACE with prior methods on the KRVQR dataset. We also report a variant of KV-TRACE that incorporates additional dense captioning information for image regions.

From Table 1, we observe that KV-TRACE substantially outperforms all baselines, including specialized knowledge-based models such as Mucko and KM-net. In particular, KV-TRACE achieves a

 doi:10.20944/preprints202511.1075.v1

Table 1. Top-1 accuracy comparisons among different models on the KRVQR dataset. All results for KV-TRACE are averaged over 5 runs.

Model	Top-1 Accuracy (%)
Q-type [7]	8.12
LSTM [7]	8.94
FiLM [23]	17.32
MFH [38]	19.87
UpDown [1]	22.14
MCAN [37]	22.81
Mucko [40]	24.26
KM-net [6]	25.43
KV-TRACE (2 steps)	**31.72**
KV-TRACE + Dense Captioning (2 steps)	**32.05**

gain of more than 6 percentage points over KM-net, highlighting the benefit of combining dynamic key–value memory with spatial-aware graph reasoning and iterative multi-hop inference. When augmented with dense captioning (KV-TRACE + Dense Captioning), the performance improves slightly further, although a paired t-test over multiple runs shows that this improvement is only marginally significant, indicating that most of the gains come from the core architecture of KV-TRACE.

4.7. Main Results on FVQA

Table 2 reports top-1 and top-3 accuracies on the FVQA dataset. We compare KV-TRACE against the FVQA ensemble model, several baselines from Narasimhan, Lazebnik, and Schwing [18], Narasimhan and Schwing [19], Mucko, and GRUC variants.

Table 2. Top-1 and top-3 accuracy of different models on the FVQA dataset. (1: Narasimhan and Schwing [19], 2: Narasimhan, Lazebnik, and Schwing [18])

Model	Accuracy (%)	
	Top-1	Top-3
FVQA (Ensemble) [32]	58.76	–
$STTF^1$	62.20	75.60
OB^2	69.58	80.47
Mucko [40]	73.21	86.03
GRUC [36]	79.81	91.30
GRUC (without Semantic graph)	78.24	87.92
KV-TRACE (1 step)	79.95	90.84
KV-TRACE + Dense Captioning (1 step)	**81.43**	**95.62**

On FVQA, KV-TRACE already matches or slightly surpasses the performance of GRUC without leveraging dense captions. When dense captioning is incorporated, KV-TRACE + Dense Captioning achieves new state-of-the-art performance, improving top-1 accuracy by approximately 1.6 percentage points and top-3 accuracy by over 4 points compared to GRUC. Interestingly, dense captioning plays a more pronounced role on FVQA than on KRVQR, which may be attributed to FVQA's bias toward single-hop reasoning, where local textual descriptions of image regions can more directly support factual grounding.

4.8. Ablation: Number of Reasoning Steps

To investigate the effect of the number of reasoning steps T in the iterative module, we vary T from 1 to 4 while keeping all other hyperparameters fixed. Table 3 summarizes the results on the KRVQR dataset.

The results indicate that the best performance is obtained with two reasoning steps, which aligns well with the dataset's annotation that includes up to two-hop questions. Using only one step leads to slightly lower performance, suggesting that a single-pass inference is insufficient to capture multi-hop

 doi:10.20944/preprints202511.1075.v1

Table 3. Top-1 accuracy of KV-TRACE on the KRVQR dataset with different numbers of reasoning steps.

Model (KV-TRACE)	Top-1 Accuracy (%)
1 step	30.89
2 steps	**31.72**
3 steps	27.36
4 steps	26.51

dependencies. In contrast, using more than two steps (three or four) causes a noticeable degradation in performance, likely due to overfitting and accumulated noise from unnecessary iterative updates when the data do not require deeper reasoning. This confirms that the iterative design of KV-TRACE is effective when properly aligned with the underlying reasoning depth of the dataset.

4.9. Ablation: Memory Module Variants

We next examine the contribution of the proposed dynamic key–value memory module by replacing it with alternative memory structures and measuring performance on KRVQR. We consider three variants: (i) a simple average-embedding memory, (ii) a standard key–value memory [17], and (iii) the proposed dynamic key–value memory. Table 4 presents the results.

Table 4. Top-1 accuracy of KV-TRACE on KRVQR with different memory module variants.

Memory Model in KV-TRACE	Top-1 Accuracy (%)
Average embedding memory module	27.68
Key–value memory (subject+relation $\rightarrow$ object)	26.91
Proposed dynamic key–value memory	**31.72**

The comparison demonstrates that the proposed dynamic key–value memory substantially outperforms the other two variants. The conventional key–value memory, which uses subject–relation pairs as keys and the object as value, falls short in this setting because many questions in KRVQR involve reasoning about different elements of a triplet (e.g., subject or relation, rather than only the object). A simple average-embedding memory performs slightly better than the standard key–value memory, as it does not impose a rigid subject–relation/object decomposition, but it still lacks the flexibility and expressivity of our dynamic triplet-level memory representation. These observations confirm the importance of designing a memory structure that can symmetrically reason about all parts of a knowledge triplet.

4.10. Ablation: Knowledge-Guided Graph Reasoning

Finally, we analyze the impact of injecting external knowledge into the visual graph reasoning module. To this end, we compare the full KV-TRACE model with a variant that disables knowledge-guided reasoning on the spatial-aware image graph, i.e., the graph attention is conditioned only on the question encoding but not on the retrieved knowledge triplets. Results are shown in Table 5.

Table 5. Effect of knowledge-guided visual graph reasoning on KRVQR. We compare the full KV-TRACE against a variant that does not use retrieved knowledge when computing visual attention.

Model Variant (KV-TRACE)	Top-1 Accuracy (%)
w/o knowledge-guided reasoning	29.98
Full model (with knowledge-guided graph)	**31.72**

We observe that removing knowledge guidance from the visual graph leads to a drop of nearly 1.8 percentage points in top-1 accuracy. This indicates that symbolic knowledge not only supports answer prediction directly but also helps the model attend to more relevant regions and relations in the image. In other words, knowledge-guided graph reasoning provides an important synergy: the retrieved

 doi:10.20944/preprints202511.1075.v1

triplets refine the visual attention patterns, and the structured visual context, in turn, disambiguates which knowledge facts are most useful for answering the question.

4.11. Additional Analyses and Discussion

Beyond the standard ablations, we also explore the effect of varying the number of retrieved facts and the maximum number of detected objects used to construct the spatial graph. Table 6 illustrates the influence of different values of k, the number of top-ranked knowledge triplets stored in the memory.

Table 6. Effect of the number of retrieved knowledge triplets k on the KRVQR dataset.

Number of retrieved facts k	Top-1 Accuracy (%)
$k = 3$	30.91
$k = 5$	**31.72**
$k = 8$	31.21
$k = 10$	30.58

The results suggest that retrieving too few facts may omit important information, whereas retrieving too many facts introduces noise that can distract the reasoning process. Empirically, using $k = 5$ retrieved facts provides a good balance between coverage and noise, which is consistent with the setting described in the methodology.

In a similar manner, we control the number of detected objects r used to build the spatial-aware graph. Using very small r can cause the model to miss relevant visual entities, while very large r increases computational cost and may lead to an overly dense or noisy graph. We find that $r = 36$ yields a favorable trade-off, echoing prior work in object-based VQA models.

4.12. Qualitative and Error Analysis

To further understand how KV-TRACE performs multi-hop reasoning, we conduct qualitative analyses on randomly selected examples from the KRVQR test set. For correctly answered cases, we observe that the model tends to (i) assign high attention scores to the fact triplets that are semantically closest to the question entities and relations, and (ii) focus visual attention on object regions that are strongly linked to these facts. For example, when asked about the biological category of an object and its relation to another object, the model first retrieves the relevant taxonomic triplet from the KB and then attends to the corresponding visual object before generating the final answer.

For incorrectly answered questions, common failure patterns include: (1) ambiguity in the visual scene, such as multiple similar objects (e.g., several persons or tools) where the model attends to the wrong instance; (2) incomplete or noisy knowledge retrieval, where the relevant fact is not among the top-k retrieved triplets; and (3) compounding errors in iterative reasoning when the initial attention is misaligned, leading later iterations to reinforce an incorrect hypothesis.

These observations suggest that future work could focus on improving ambiguity resolution in crowded scenes, designing more robust knowledge retrieval mechanisms, and exploring uncertainty-aware iterative reasoning strategies that can detect and correct misaligned intermediate steps rather than amplifying them.

Overall, the experimental results demonstrate that KV-TRACE consistently outperforms strong baselines across two benchmarks, while the detailed ablations confirm the importance of each architectural component, including the dynamic key–value memory, knowledge-guided graph reasoning, and the iterative multi-hop inference module.

5. Conclusions

In this work, we introduced KV-TRACE, a comprehensive multi-step reasoning framework that unifies explicit symbolic inference and implicit visual relational reasoning through a dynamically structured key-value memory and a spatially grounded graph-based visual encoder. Our approach departs from conventional knowledge-based VQA pipelines by tightly integrating knowledge retrieval,

 doi:10.20944/preprints202511.1075.v1

memory-augmented reasoning, and graph-structured visual understanding into a single, iterative architecture capable of refining its internal representations over multiple reasoning cycles.

Across the reasoning process, KV-TRACE first performs explicit knowledge grounding by addressing relevant triplets stored in a dynamic memory bank, where each key-value slot encodes the semantic components of an RDF fact. This symbolic inference is then complemented by implicit visual graph reasoning, where spatially related objects and their relational cues are propagated through multi-head graph attention, enabling the model to incorporate both conceptual knowledge and visual regularities. The iterative design ensures that these two streams of reasoning—linguistic-symbolic and relational-visual—reinforce one another, allowing the model to progressively converge toward a coherent, knowledge-consistent interpretation of the question and image.

Extensive experiments on two challenging knowledge-based visual question answering benchmarks, KRVQR and FVQA, demonstrate the effectiveness of the proposed architecture. KV-TRACE achieves new state-of-the-art results on both datasets, benefiting from its ability to reason over multi-hop knowledge chains and spatial configurations in complex scenes. The improvements are consistent across one-step and two-step reasoning questions, indicating that the design is inherently flexible and well-suited for tasks requiring diverse types of inferential behavior, ranging from attribute lookup to relational chaining to cross-modal grounding.

Furthermore, the modular design of KV-TRACE opens several promising research directions. For instance, the dynamic memory component could be extended to handle richer forms of structured knowledge, including hierarchical knowledge graphs, probabilistic rules, or large-scale encyclopedic corpora. Similarly, the graph reasoning module could benefit from more advanced geometric or causal relational modeling to capture fine-grained scene dynamics or deeper causal dependencies. Another fruitful direction lies in exploring reinforcement learning–based training schemes that allow the model to autonomously select reasoning strategies, thereby enabling adaptive multi-hop inference beyond fixed-step architectures.

In summary, KV-TRACE offers a unified, interpretable, and extensible framework for knowledge-grounded multimodal reasoning. By integrating structured symbolic memories with spatially aware graph attention, it provides a principled approach toward bridging visual perception and knowledge-based inference. We believe that this work not only contributes a strong method for KVQA but also lays the foundation for future progress in multimodal reasoning systems that require richer forms of symbolic–visual integration, such as embodied QA, scientific diagram understanding, and real-world decision-making agents.

References

1. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 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*, 6077–6086.
2. Andreas, J.; Rohrbach, M.; Darrell, T.; and Klein, D. 2016. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 39–48.
3. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2425–2433.
4. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. 2007. DBpedia: A nucleus for a web of open data. In *The Semantic Web*, 722–735. Springer.
5. Ben-younes, H.; Cadene, R.; Cord, M.; and Thome, N. 2017. MUTAN: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
6. Cao, Q.; Li, B.; Liang, X.; and Lin, L. 2019. Explainable high-order visual question reasoning: A new benchmark and knowledge-routed network. *arXiv preprint arXiv:1909.10128*.
7. Cao, Q.; Li, B.; Liang, X.; Wang, K.; and Lin, L. 2021. Knowledge-routed visual question reasoning: Challenges for deep representation embedding. *IEEE Transactions on Neural Networks and Learning Systems*.
8. Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

 doi:10.20944/preprints202511.1075.v1

9. Gu, J.; Zhao, H.; Lin, Z.; Li, S.; Cai, J.; and Ling, M. 2019. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1969–1978.
10. Hu, R.; Rohrbach, A.; Darrell, T.; and Saenko, K. 2019. Language-conditioned graph networks for relational reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10294–10303.
11. Hudson, D. A.; and Manning, C. D. 2019. GQA: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506*, 3(8).
12. Jain, A.; Kothyari, M.; Kumar, V.; Jyothi, P.; Ramakrishnan, G.; and Chakrabarti, S. 2021. Select, substitute, search: A new benchmark for knowledge-augmented visual question answering. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.
13. Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
14. Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. *Advances in Neural Information Processing Systems*, 29: 289–297.
15. Ma, L.; Lu, Z.; and Li, H. 2016. Learning to answer questions from image using convolutional neural network. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*.
16. Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3195–3204.
17. Miller, A.; Fisch, A.; Dodge, J.; Karimi, A.-H.; Bordes, A.; and Weston, J. 2016. Key-value memory networks for directly reading documents. *arXiv preprint arXiv:1606.03126*.
18. Narasimhan, M.; Lazebnik, S.; and Schwing, A. G. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. *arXiv preprint arXiv:1811.00538*.
19. Narasimhan, M.; and Schwing, A. G. 2018. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. 451–468.
20. Norcliffe-Brown, W.; Vafeias, E.; and Parisot, S. 2018. Learning conditioned graph structures for interpretable visual question answering. *arXiv preprint arXiv:1806.07243*.
21. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32: 8026–8037.
22. Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.
23. Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
24. Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; and Manning, C. D. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.
25. Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28: 91–99.
26. Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-first AAAI Conference on Artificial Intelligence*.
27. Sukhbaatar, S.; Szlam, A.; Weston, J.; and Fergus, R. 2015. End-to-end memory networks. *arXiv preprint arXiv:1503.08895*.
28. Tandon, N.; De Melo, G.; Suchanek, F.; and Weikum, G. 2014. Webchild: Harvesting and organizing commonsense knowledge from the Web. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, 523–532.
29. Teney, D.; Liu, L.; and van den Hengel, A. 2017. Graph-structured representations for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.
30. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2017. Graph attention networks. *6th International Conference on Learning Representations*.
31. Wang, P.; Wu, Q.; Cao, J.; Shen, C.; Gao, L.; and van den Hengel, A. 2019. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1960–1968.
32. Wang, P.; Wu, Q.; Shen, C.; Dick, A.; and van den Hengel, A. 2017. FVQA: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10): 2413–2427.
33. Weston, J.; Chopra, S.; and Bordes, A. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.

34. Xu, K.; Lai, Y.; Feng, Y.; and Wang, Z. 2019. Enhancing key-value memory neural networks for knowledge based question answering. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2937–2947.
35. Yang, Z.; He, X.; Gao, J.; Deng, L.; and Smola, A. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 21–29.
36. Yu, J.; Zhu, Z.; Wang, Y.; Zhang, W.; Hu, Y.; and Tan, J. 2020. Cross-modal knowledge reasoning for knowledge-based visual question answering. *Pattern Recognition*, 108: 107563.
37. Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; and Tian, Q. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6281–6290.
38. Yu, Z.; Yu, J.; Xiang, C.; Fan, J.; and Tao, D. 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 29(12): 5947–5959.
39. Zareian, A.; Karaman, S.; and Chang, S.-F. 2020. Bridging knowledge graphs to generate scene graphs. In *Proceedings of the European Conference on Computer Vision*, 606–623. Springer.
40. Zhu, Z.; Yu, J.; Wang, Y.; Sun, Y.; Hu, Y.; and Wu, Q. 2020. Mucko: Multi-Layer Cross-Modal Knowledge Reasoning for Fact-based Visual Question Answering. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 1097–1103. International Joint Conferences on Artificial Intelligence Organization.
41. Ziaeefard, M.; and Lécué, F. 2020. Towards knowledge-augmented visual question answering. In *Proceedings of the 28th International Conference on Computational Linguistics*, 1863–1873.
42. Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, 382–398. Springer.
43. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 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*, 6077–6086.
44. Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations*.
45. Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.
46. Chen, H.; Ding, G.; Zhao, S.; and Han, J. 2018. Temporal-difference learning with sampling baseline for image captioning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
47. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; and Chua, T.-S. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6298–6306. IEEE.
48. Elliott, D.; and Keller, F. 2013. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1292–1302.
49. Erden, M. S.; and Tomiyama, T. 2010. Human-Intent Detection and Physically Interactive Control of a Robot Without Force Sensors. *IEEE Transactions on Robotics* 26(2): 370–382.
50. Fang, H.; Gupta, S.; Iandola, F.; Srivastava, R. K.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J. C.; et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1473–1482.
51. Gao, J.; Wang, S.; Wang, S.; Ma, S.; and Gao, W. 2019. Self-critical n-step Training for Image Captioning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* .
52. Guo, L.; Liu, J.; Lu, S.; and Lu, H. 2019. Show, tell and polish: Ruminant decoding for image captioning. *IEEE Transactions on Multimedia* .
53. Herdade, S.; Kappeler, A.; Boakye, K.; and Soares, J. 2019. Image captioning: Transforming objects into words. In *Advances in Neural Information Processing Systems*, 11137–11147.
54. Huang, L.; Wang, W.; Chen, J.; and Wei, X.-Y. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, 4634–4643.
55. Karpathy, A.; Joulin, A.; and Li, F. F. 2014. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. *Advances in neural information processing systems* 3.
56. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123(1): 32–73.

 doi:10.20944/preprints202511.1075.v1

57. Kuznetsova, P.; Ordonez, V.; Berg, A. C.; Berg, T. L.; and Choi, Y. 2012. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, 359–368. Association for Computational Linguistics.
58. Li, G.; Zhu, L.; Liu, P.; and Yang, Y. 2019. Entangled transformer for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, 8928–8937.
59. Li, S.; Kulkarni, G.; Berg, T. L.; Berg, A. C.; and Choi, Y. 2011. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, 220–228. Association for Computational Linguistics.
60. Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.
61. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.
62. Liu, D.; Zha, Z.-J.; Zhang, H.; Zhang, Y.; and Wu, F. 2018. Context-aware visual policy network for sequence-level image captioning. *Proceedings of the 26th ACM international conference on Multimedia* .
63. Liu, S.; Zhu, Z.; Ye, N.; Guadarrama, S.; and Murphy, K. 2017. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, 873–881.
64. Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 375–383.
65. Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2018. Neural Baby Talk. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
66. Mitchell, M.; Han, X.; Dodge, J.; Mensch, A.; Goyal, A.; Berg, A.; Yamaguchi, K.; Berg, T.; Stratos, K.; and Daumé III, H. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 747–756. Association for Computational Linguistics.
67. Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318.
68. Qin, Y.; Du, J.; Zhang, Y.; and Lu, H. 2019. Look Back and Predict Forward in Image Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8367–8375.
69. Ranzato, M.; Chopra, S.; Auli, M.; and Zaremba, W. 2015. Sequence Level Training with Recurrent Neural Networks. *International Conference on Learning Representations* .
70. Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7008–7024.
71. Schmidt, P.; Mael, E.; and Wurtz, R. P. 2006. A sensor for dynamic tactile information with applications in human-robot interaction and object exploration. *Robotics and Autonomous Systems* 54(12): 1005–1014.
72. Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.
73. Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3156–3164.
74. Vo, N.; Jiang, L.; Sun, C.; Murphy, K.; Li, L.; Feifei, L.; and Hays, J. 2019. Composing Text and Image for Image Retrieval - an Empirical Odyssey 6439–6448.
75. Wang, L.; Li, Y.; Huang, J.; and Lazebnik, S. 2019. Learning Two-Branch Neural Networks for Image-Text Matching Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(2): 394–407.
76. Wang, L.; Schwing, A.; and Lazebnik, S. 2017. Diverse and Accurate Image Description Using a Variational Auto-Encoder with an Additive Gaussian Encoding Space. In *Advances in Neural Information Processing Systems 30*, 5756–5766.
77. Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8(3-4): 229–256.
78. Williams, R. J.; and Zipser, D. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation* 1(2): 270–280.
79. Wu, Q.; Wang, P.; Shen, C.; Reid, I.; and Hengel, A. 2018. Are You Talking to Me? Reasoned Visual Dialog Generation Through Adversarial Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6106–6115.

 doi:10.20944/preprints202511.1075.v1

80. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, 2048–2057.
81. Yang, X.; Tang, K.; Zhang, H.; and Cai, J. 2019. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10685–10694.
82. Yang, Z.; Yuan, Y.; Wu, Y.; Cohen, W. W.; and Salakhutdinov, R. R. 2016. Review Networks for Caption Generation. In *Advances in Neural Information Processing Systems 29*, 2361–2369.
83. Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2018. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, 684–699.
84. Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2019. Hierarchy Parsing for Image Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
85. You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image Captioning With Semantic Attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
86. Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.
87. Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.
88. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186.
89. Magdalena Kaiser, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Reinforcement Learning from Reformulations In Conversational Question Answering over Knowledge Graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 459–469.
90. Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 4483–4491. Survey Track.
91. Yunshi Lan and Jing Jiang. 2021. Modeling transitions of focal entities for conversational knowledge base question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.
92. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
93. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
94. Meishan Zhang, Hao Fei, Bin Wang, Shengqiong Wu, Yixin Cao, Fei Li, and Min Zhang. Recognizing everything from all modalities at once: Grounded multimodal universal information extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.
95. Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.
96. Shengqiong Wu, Hao Fei, Jingkang Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Tat-seng Chua. Learning 4d panoptic scene graph generation from rich 2d visual scene. *Proceedings of the CVPR*, 2025.
97. Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.
98. Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, Qingyu Shi, Zhiyuan Zhou, Liangtao Shi, Minghe Gao, Daoan Zhang, Zhiqi Ge, Weiming Wu, Siliang Tang, Kaihang Pan, Yaobo Ye, Haobo Yuan, Tao Zhang, Tianjie Ju, Zixiang Meng, Shilin Xu, Liyu Jia, Wentao Hu, Meng Luo, Jiebo Luo, Tat-Seng Chua, Shuicheng Yan, and Hanwang Zhang. On path to multimodal generalist: General-level and general-bench. In *Proceedings of the ICML*, 2025.
99. Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024.

 doi:10.20944/preprints202511.1075.v1

100. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. URL http://dx.doi.org/10.1038/nature14539.
101. Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014. URL https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.
102. Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016. URL http://arxiv.org/abs/1604.08608.
103. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.
104. Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
105. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
106. Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.
107. Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
108. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
109. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
110. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
111. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
112. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
113. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.
114. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
115. Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.
116. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. *arXiv preprint arXiv:2305.11719*, 2023.
117. Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*, 2024.
118. Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.
119. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
120. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.

 doi:10.20944/preprints202511.1075.v1

121. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
122. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
123. Bobo Li, Hao Fei, Fei Li, Tat-seng Chua, and Donghong Ji. 2024. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Transactions on Multimedia Computing, Communications and Applications* (2024).
124. Bobo Li, Hao Fei, Fei Li, Shengqiong Wu, Lizi Liao, Yinwei Wei, Tat-Seng Chua, and Donghong Ji. 2025. Revisiting conversation discourse for dialogue disentanglement. *ACM Transactions on Information Systems* 43, 1 (2025), 1–34.
125. Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. 2023. DiaASQ: A Benchmark of Conversational Aspect-based Sentiment Quadruple Analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*. 13449–13467.
126. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Fangfang Su, Fei Li, and Donghong Ji. 2024. Harnessing holistic discourse features and triadic interaction for sentiment quadruple extraction in dialogues. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 18462–18470.
127. Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. 2025. Combating Multimodal LLM Hallucination via Bottom-Up Holistic Reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 8460–8468.
128. Shengqiong Wu, Weicai Ye, Jiahao Wang, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Shuicheng Yan, Hao Fei, et al. 2025. Any2caption: Interpreting any condition to caption for controllable video generation. *arXiv preprint arXiv:2503.24379* (2025).
129. Han Zhang, Zixiang Meng, Meng Luo, Hong Han, Lizi Liao, Erik Cambria, and Hao Fei. 2025. Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark. In *Proceedings of the ACM on Web Conference 2025*. 2872–2881.
130. Hao Fei, Yafeng Ren, and Donghong Ji. 2020, A tree-based neural network model for biomedical event trigger detection, Information Sciences, 512, 175
131. Hao Fei, Yafeng Ren, and Donghong Ji. 2020, Dispatched attention with multi-task learning for nested mention recognition, Information Sciences, 513, 241
132. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. 2021, A span-graph neural model for overlapping entity relation extraction in biomedical texts, Bioinformatics, 37, 1581
133. Yu Zhao, Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, and Tat-seng Chua. 2025. Grammar induction from visual, speech and text. *Artificial Intelligence* 341 (2025), 104306.
134. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
135. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12794–12802, 2021.
136. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
137. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, 2021.
138. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
139. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*, 2023.
140. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/N19-1423.
141. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
142. Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

 doi:10.20944/preprints202511.1075.v1

143. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024.
144. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
145. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7641–7653, 2024.
146. Mihir Momaya, Anjnya Khanna, Jessica Sadavarte, and Manoj Sankhe. Krushi–the farmer chatbot. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–6. IEEE, 2021.
147. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.
148. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
149. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
150. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
151. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
152. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2024,*, 2024.
153. Abbott Chen and Chai Liu. Intelligent commerce facilitates education technology: The platform and chatbot for the taiwan agriculture service. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 11:1–10, 01 2021.
154. Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024.
155. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
156. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.
157. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.
158. Isakwisa Gaddy Tende, Kentaro Aburada, Hisaaki Yamaba, Tetsuro Katayama, and Naonobu Okazaki. Proposal for a crop protection information system for rural farmers in tanzania. *Agronomy*, 11(12):2411, 2021.
159. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
160. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
161. Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. Farmchat: a conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–22, 2018.
162. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 79240–79259, 2023.

 doi:10.20944/preprints202511.1075.v1

163. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
164. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2):50:1–50:32, 2023.
165. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023.
166. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.
167. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5544–5556, 2023.
168. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.
169. Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.
170. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.