Differentiable Retrieval-Guided Multimodal Reasoning for Knowledge-Intensive Visual Question Understanding

Camille Dupuis *, Juliette Declerck, Elodie Fairchild, Thomas Damme

Posted Date: 23 October 2025

doi: 10.20944/preprints202510.1820.v1

Keywords: knowledge-grounded vision-language reasoning; visual question answering; retrieval-augmented generation; multimodal learning; commonsense knowledge integration; external knowledge retrieval

 doi:10.20944/preprints202510.1820.v1

Article

Differentiable Retrieval-Guided Multimodal Reasoning for Knowledge-Intensive Visual Question Understanding

Camille Dupuis *, Juliette Declerck, Elodie Fairchild and Thomas Damme

Université libre de Bruxelles
* Correspondence: camille.dupuis@ulb.be

Abstract

Visual understanding in real-world scenarios often extends far beyond what is directly visible in an image, requiring the ability to reason with external and commonsense knowledge. Traditional Visual Question Answering (VQA) systems, while powerful in multimodal comprehension, typically confine their reasoning to the visual scene, making them inadequate when contextual or encyclopedic information is essential for accurate answers. To address this limitation, we introduce **KnowSight**, a unified framework for *knowledge-grounded visual question answering*, which integrates retrieval-based external knowledge reasoning with multimodal understanding in a single end-to-end architecture. Unlike prior works that separate document retrieval and answer generation, KnowSight establishes a joint optimization scheme that enables the model to dynamically align visual semantics with relevant knowledge sources. Our design incorporates a differentiable retrieval process that allows backpropagation through document scoring, ensuring that knowledge selection is directly informed by the downstream reasoning objective. This paradigm bridges the gap between perception and cognition, allowing the model to answer questions that require factual grounding, causal inference, or commonsense understanding. Comprehensive experiments on OK-VQA and related benchmarks demonstrate that KnowSight significantly s urpasses p revious r etrieval-augmented s ystems i n b oth knowledge efficiency and i nterpretability. Furthermore, we propose a new set of diagnostic metrics to disentangle the contributions of visual grounding and knowledge retrieval. Our analysis reveals that integrating structured and unstructured knowledge through joint training substantially reduces reliance on large retrieval sets, leading to faster convergence and more robust reasoning performance. Beyond outperforming existing methods, KnowSight offers a generalized blueprint for multimodal reasoning systems that continuously learn and adapt their external knowledge grounding in open-world environments.

Keywords: knowledge-grounded vision-language reasoning; visual question answering; retrieval-augmented generation; multimodal learning; commonsense knowledge integration; external knowledge retrieval

1. Introduction

Visual Question Answering (VQA) has emerged as a fundamental problem at the intersection of Computer Vision, Natural Language Processing, and Knowledge Representation. The task requires an intelligent system to generate a correct and contextually grounded answer to a natural-language question about an image. Early VQA models primarily focused on direct visual-textual alignment, leveraging deep multimodal encoders to integrate features from the visual and linguistic modalities. However, as the field matured, it became evident that many questions posed to visual systems require reasoning that extends far beyond perceptual content alone.

In conventional datasets, such as VQA-v2, the visual evidence is often sufficient to determine the correct answer. In contrast, knowledge-based VQA (KB-VQA) and Outside-Knowledge VQA

 doi:10.20944/preprints202510.1820.v1

(OK-VQA) [29] push the boundaries of this paradigm by introducing questions that require external factual or commonsense information. For instance, answering a question like "Why is the man holding an umbrella on a sunny day?" demands background understanding of cultural practices or contextual reasoning about potential scenarios, rather than visual pattern recognition alone. This highlights a critical limitation in current multimodal systems: their dependency on the information available within the image-text pair, without access to external repositories of human knowledge.

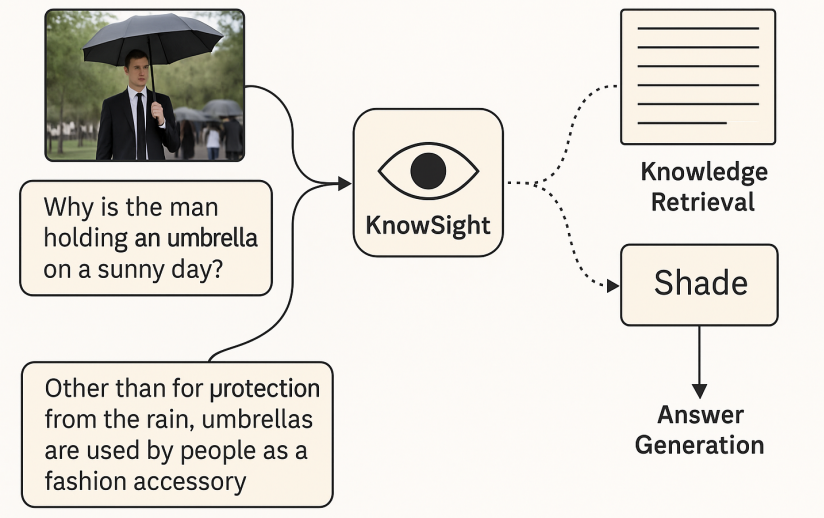

Figure 1. The motivation of this work.

To address this, recent works have turned to retrieval-based augmentation, wherein external documents or facts are retrieved from large knowledge bases such as Wikipedia [8,26,38]. Dense Passage Retrieval (DPR) [17] has proven particularly effective for mapping questions and documents into shared vector spaces, allowing for semantic retrieval of relevant passages. However, existing systems typically train the retriever and the answer generator independently, creating a disconnect between knowledge selection and downstream reasoning. This leads to inefficiencies: retrievers may prioritize documents that merely contain surface-level overlaps with query terms, while generators must process redundant or misleading content. Prior research attempted to alleviate this issue through heuristic pseudo-labels [32], yet such approaches introduce noisy supervision and may not reflect true semantic relevance.

Decoupling retrieval and generation causes several systemic weaknesses. Firstly, the retriever cannot be optimized to improve the answer generation objective directly. Secondly, the generator may become over-reliant on spurious correlations between retrieved content and ground-truth answers. Finally, such pipelines require a large number of retrieved documents (often 50 or more [8,10]), resulting in substantial computational overhead and potential degradation of answer quality. These factors motivate the need for an end-to-end differentiable framework that can jointly refine both retrieval and reasoning.

Retrieval-Augmented Generation (RAG) [21] first demonstrated that joint optimization between a retriever and generator could improve open-domain QA performance. Yet, when directly applied to multimodal settings such as OK-VQA, RAG exhibits critical weaknesses. Our preliminary studies confirmed that RAG's loss formulation often misattributes credit to irrelevant documents when visual cues alone suffice for question answering. Moreover, many OK-VQA samples require compositional reasoning between visual context and external text—a challenge RAG was not designed to handle. Consequently, naive adaptations of RAG yield inconsistent performance across varying question types and fail to leverage the full potential of visual grounding.

Motivated by these limitations, we propose **KnowSight**, a *retrieval-in-the-loop reasoning framework* that explicitly integrates knowledge retrieval into the visual question answering process. KnowSight's central innovation lies in its unified objective function, which couples retrieval confidence with genera-

 doi:10.20944/preprints202510.1820.v1

tive correctness, ensuring that documents contributing to accurate reasoning are reinforced during training. This approach eliminates the need for heuristic pseudo-relevance labels and enables dynamic feedback between retrieval and generation modules. The retriever learns to focus on documents that truly aid reasoning, while the generator refines its textual grounding on verified contextual evidence.

Beyond its architectural novelty, KnowSight introduces a diagnostic methodology for analyzing retrieval–generation interactions, measuring both relevance precision and reasoning attribution. Experiments on OK-VQA and extended benchmarks show that KnowSight consistently outperforms traditional two-step models and RAG-style systems in accuracy, efficiency, and knowledge utilization. It requires significantly fewer retrieved documents—often as few as 5 to 10 per question—without sacrificing quality. This efficiency makes it practical for large-scale multimodal applications and continuous learning scenarios.

The implications of KnowSight extend beyond OK-VQA. As multimodal large language models (MLLMs) increasingly shape the landscape of AI, the ability to ground reasoning on dynamically retrievable knowledge will be crucial for building interpretable, reliable, and adaptive systems. KnowSight offers a generalizable framework that could be extended to visual dialogue, multimodal commonsense inference, and even agentic reasoning systems capable of autonomous knowledge acquisition. Ultimately, this work moves towards a long-term vision of *open-world multimodal understanding*, where models can see, reason, and learn continuously by drawing from an ever-evolving body of human knowledge.

In summary, our contributions are threefold:

- We introduce **KnowSight**, a unified joint-learning framework for retrieval-augmented visual reasoning that integrates external knowledge dynamically during answer generation.
- We propose a diagnostic perspective on retrieval–generation synergy, offering fine-grained insights into how knowledge retrieval influences multimodal reasoning outcomes.
- We demonstrate state-of-the-art performance on OK-VQA and related benchmarks, achieving significant efficiency improvements in both retrieval quality and computational cost.

This work thus bridges a key gap between perception and knowledge, contributing to the broader agenda of developing multimodal systems capable of grounded, explainable, and knowledge-driven reasoning in complex visual environments.

2. Related Work

2.1. Open-Domain Question Answering Paradigms

Open-domain Question Answering (QA) represents a long-standing research problem at the intersection of information retrieval and natural language understanding. The core objective is to enable models to answer arbitrary natural language questions by accessing diverse sources of world knowledge. In early QA systems, knowledge was primarily derived from static textual corpora or manually curated knowledge bases. However, recent progress in large-scale pre-training [3,12,21,35] and differentiable retrieval frameworks [13,20] has shifted the paradigm toward dynamic, context-aware retrieval augmentation.

Parametric QA models such as T5 [35] embed a vast amount of factual knowledge within model parameters, allowing "closed-book" inference. Yet, their knowledge coverage remains bounded by the training corpus, making them less reliable for long-tail facts. In contrast, retrieval-augmented approaches explicitly query large document stores or web-scale knowledge repositories to extract supporting evidence dynamically. This design paradigm increases interpretability and adaptability while maintaining scalability. Pioneering works such as REALM [12] and ORQA [20] integrated neural retrievers with downstream QA objectives, establishing differentiable retrieval as a key milestone. Later, RAG [21] extended this idea into a generative setting, allowing the model to produce free-form textual answers while marginalizing over retrieved passages.

Despite their impressive success, these systems often face inherent trade-offs between retrieval precision, model interpretability, and computational overhead. Subsequent research has aimed to

 doi:10.20944/preprints202510.1820.v1

refine retriever–generator synergy by jointly training both components and improving the quality of contextualized representations [6,27,36]. Our work builds on this line of research by extending differentiable retrieval to multimodal contexts, where visual semantics must guide the selection of textual evidence.

2.2. Evolution of Multimodal Visual Question Answering

Visual Question Answering (VQA) tasks require systems to reason jointly over image content and textual queries. Early VQA methods relied on simple concatenation or attention-based fusion of CNN and RNN representations [15,37,47], which limited their capacity for complex cross-modal reasoning. With the advent of large-scale multimodal pre-training, transformer-based architectures [4,14,24,39,42, 48] became dominant, unifying visual and textual representations within shared embedding spaces.

Recent advancements such as UNITER [4] and VinVL [48] demonstrate that learning fine-grained object–text alignments through large-scale pre-training significantly enhances zero-shot and few-shot generalization. Moreover, the emergence of multimodal LLMs (e.g., Flamingo, BLIP-2) has shown that injecting structured knowledge into vision-language reasoning pipelines improves factual grounding and commonsense understanding. Yet, most conventional VQA models are "closed-world" learners—they rely solely on image-text data seen during training and lack the capacity to access evolving external knowledge.

2.3. Knowledge-Based VQA and External Knowledge Integration

Knowledge-based VQA (KB-VQA) extends traditional VQA by incorporating structured and unstructured knowledge resources to enrich reasoning. Structured resources include ConceptNet [38], Freebase, or other Knowledge Graphs (KGs), which provide explicit relations between entities. Unstructured sources such as Wikipedia or the web offer contextual evidence and open-domain factual grounding. Hybrid frameworks [8,9,28,44] attempt to bridge these modalities by aligning symbolic and textual knowledge representations.

ConceptBERT [9] first demonstrated that embedding graph nodes into a transformer attention framework enables semantic alignment between entities and visual elements. KRISP [28] extended this idea with a symbolic knowledge module that links image regions and question tokens with KG entities, enhancing interpretability. MAVEx [44] further broadened the scope by leveraging multiple heterogeneous sources (Google Images, ConceptNet, and Wikipedia) to validate candidate answers. Meanwhile, TRiG [8] and KAT [10] integrated dense retrieval and large language models (T5, GPT-3) into VQA, showcasing the benefits of language-grounded retrieval for multimodal reasoning.

Despite these efforts, current KB-VQA systems often treat knowledge retrieval and reasoning as disjoint stages. This leads to suboptimal alignment between retrieved evidence and the generative reasoning process. Moreover, reliance on pseudo-relevance labels [26,32] can introduce noise, as retrieved passages containing the answer text are not guaranteed to be semantically relevant. Our approach addresses these challenges through a joint optimization paradigm that integrates retrieval relevance directly into the reasoning objective.

2.4. Retrieval-Augmented and Differentiable Learning Frameworks

The concept of retrieval augmentation represents a pivotal shift from static knowledge embedding to dynamic reasoning architectures. Unlike fixed-parameter LMs, retrieval-augmented frameworks dynamically update their knowledge context during inference, improving adaptability to new facts without retraining. Differentiable retrieval techniques, as in REALM [12], enable backpropagation through the retrieval step, aligning document selection with downstream objectives. Building upon these advances, KnowSight introduces a multimodal differentiable retrieval mechanism that conditions knowledge selection on both linguistic and visual cues. The retrieved evidence is then used to guide answer generation, and gradients are propagated to refine the retriever's parameters based on the generation loss, achieving true end-to-end optimization.

 doi:10.20944/preprints202510.1820.v1

2.5. Recent Multimodal Reasoning and Foundation Models

With the advent of multimodal foundation models, large-scale pre-training has begun to bridge the gap between vision, language, and external knowledge. Unified architectures such as BLIP-2, OFA, and Flamingo have introduced adapter-based visual reasoning frameworks that extend LLMs' text-based commonsense reasoning to visual inputs. However, these models often rely on fixed pre-training corpora and cannot efficiently update their knowledge when encountering novel scenarios. The integration of retrieval-based reasoning mechanisms—such as those explored in KnowSight—provides a pathway toward lifelong multimodal learning, enabling dynamic access to continuously evolving world knowledge.

In addition, recent efforts explore integrating symbolic reasoning modules and neural retrieval components to improve factual consistency and interpretability. For instance, approaches leveraging neuro-symbolic alignment [30] or graph-aware prompting structures [22,44] have shown potential for mitigating hallucination in multimodal systems. Nevertheless, most methods remain limited by their inability to propagate reasoning signals back into retrieval decisions, a gap that KnowSight explicitly addresses through joint gradient optimization.

In summary, our work builds upon three converging research lines: (1) retrieval-augmented open-domain QA [12,21], (2) multimodal VQA [4,24,39], and (3) knowledge-based reasoning with structured and unstructured resources [9,28,44]. While each line contributes unique strengths, their integration into a single unified framework remains underexplored. By coupling visual-semantic encoding with differentiable retrieval and end-to-end generative reasoning, our model introduces a principled framework for *knowledge-grounded visual understanding*. This closes a critical gap between perception and reasoning, providing the foundation for a new generation of interpretable and adaptive multimodal systems.

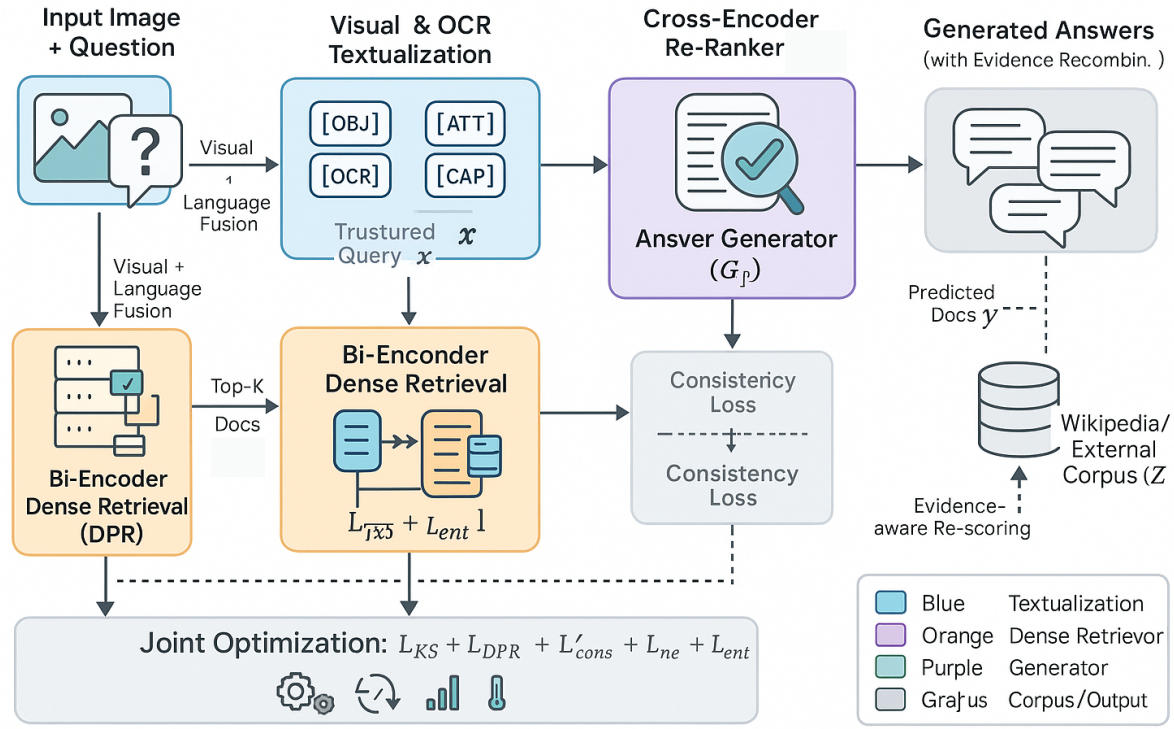

Figure 2. Overview of the proposed KnowSight framework for knowledge-intensive visual question answering. The system transforms multimodal inputs—an image and question—into a structured textual query through Visual & OCR Textualization, enabling a Bi-Encoder Dense Retriever to fetch relevant knowledge passages from an external corpus. A Cross-Encoder Re-Ranker refines the retrieved results, which are then fed into an Answer Generator for evidence-grounded prediction. The entire pipeline is optimized jointly under multiple objectives, including retrieval calibration, consistency alignment, and entropy regularization, ensuring end-to-end coupling between retrieval relevance and generative answerability.

 doi:10.20944/preprints202510.1820.v1

3. Methodology

We introduce **KnowSight**, a retrieval-in-the-loop framework for knowledge-intensive Visual Question Answering that couples multimodal query formation, weakly-supervised dense retrieval, and end-to-end joint optimization of retrieval and generation. Compared with prior RA-VQA designs, KnowSight extends (i) the vision→language reframing to include richer structured textualized cues, (ii) the retriever with calibration, re-ranking, and uncertainty-aware scoring, and (iii) the learning objective with auxiliary consistency and contrastive terms that align document relevance with answerability.

3.1. Problem Setup and Notation

Let an image–question pair be (I, q) with a set of human answers $\mathcal{S} = \{s_i\}_{i=1}^{m}$. An external corpus is $\mathcal{Z} = \{z_j\}_{j=1}^{N^d}$ (e.g., Wikipedia passages). KnowSight forms a textual query x from (I, q) (Sec. 3.2); a query encoder $\mathcal{F}_q$ and a document encoder $\mathcal{F}_d$ (both Transformer-like) map x and z to $\mathbb{R}^h$:

$$\mathbf{q} = \mathcal{F}_q(x) \in \mathbb{R}^h, \qquad \mathbf{d} = \mathcal{F}_d(z) \in \mathbb{R}^h. \tag{1}$$

A base similarity score is $r(x, z) = \mathbf{q}^\top \mathbf{d}$. Given a retrieved set $\{z_k\}_{k=1}^{K}$, a generator $\mathcal{G}_\phi$ (e.g., T5) produces an answer y conditioned on (x, z_k). KnowSight learns parameters θ (retriever) and ϕ (generator) jointly.

3.2. Multimodal-to-Text Reframing

Prior work shows that language-only Transformers can be repurposed for VQA once images are transformed into textual evidence [26,45]. KnowSight follows this paradigm but enriches the textualization with object-attribute phrases, relations, OCR strings, and a global caption.

Object and Attribute Serialization.

Using VinVL [48], we detect objects $\{o_i\}$ and attributes $\{a_{i,j}\}$ with detection and attribute confidences $\tau_o = 0.8$ and $\tau_a = 0.6$, respectively. We serialize them into normalized phrases `[OBJ]` o_i `[ATT]` $a_{i,1}, \ldots, a_{i,J_i}$ `[/OBJ]`.

Caption and OCR.

A caption c is generated by Oscar+ [48]. OCR strings $\{t_\ell\}$ are extracted using a production OCR system and normalized with case-folding and Unicode canonicalization.

Full Query String.

We concatenate all sources with type delimiters:

$$x = \texttt{[Q]}\ q\ \texttt{[/Q]} \parallel \texttt{[CAP]}\ c\ \texttt{[/CAP]} \parallel \Big\|_i\ \texttt{[REG]}\ o_i{:}\{a_{i,*}\}\ \texttt{[/REG]} \parallel \Big\|_\ell\ \texttt{[OCR]}\ t_\ell\ \texttt{[/OCR]}. \tag{2}$$

This yields a purely textual input that preserves visual grounding through structured markers and improves retrieval conditioning.

3.3. Weakly-Supervised Dense Retrieval with Calibration

KnowSight adopts Dense Passage Retrieval (DPR) with in-batch negatives [17]. The base score is $r(x, z) = \mathbf{q}^\top \mathbf{d}$, and we define a temperature-scaled retrieval probability:

$$p_\theta(z \mid x) = \frac{\exp(\lambda r(x, z))}{\sum_{j \in \mathcal{C}(x)} \exp(\lambda r(x, z_j))}, \quad \lambda > 0, \tag{3}$$

where $\mathcal{C}(x)$ is the candidate pool (top-K from FAISS or an in-batch set).

Pseudo-Relevance and Weak Labels.

 doi:10.20944/preprints202510.1820.v1

Given answer set $\mathcal{S}$, we use a pseudo relevance indicator $H(z, \mathcal{S}) = \mathbb{I}[\exists s \in \mathcal{S} \text{ s.t. } s \subset z]$ via robust string match. For each $(x, \mathcal{S})$ we pick one positive $z^+(x)$ with $H(z^+, \mathcal{S}) = 1$; other batch documents serve as negatives $\mathcal{N}(x, \mathcal{S})$.

DPR Loss.

$$\mathcal{L}_{\text{DPR}} = -\sum_{(x,\mathcal{S})\in\mathcal{T}} \log \frac{\exp(\hat{r}^+(x))}{\exp(\hat{r}^+(x)) + \sum_{z\in\mathcal{N}(x,\mathcal{S})} \exp(\hat{r}(x,z))}. \tag{4}$$

Score Calibration and Entropy Control.

To prevent overconfident early retrieval, we add an entropy-promoting regularizer on $p_\theta(\cdot \mid x)$:

$$\mathcal{L}_{\text{ent}} = \sum_x - \sum_{z\in\mathcal{C}(x)} p_\theta(z \mid x) \log p_\theta(z \mid x), \tag{5}$$

and optimize λ by backpropagation. This encourages a gentler distribution over candidates early in training and sharpens as learning progresses.

3.4. Cross-Encoder Consistency Re-Ranking

Beyond bi-encoder DPR, KnowSight employs a light cross-encoder $\mathcal{R}_\psi$ that scores (x, z) jointly:

$$s_{\text{ce}}(x, z) = \mathcal{R}_\psi(\texttt{[CLS]}\ x\ \texttt{[SEP]}\ z\ \texttt{[CLS]}). \tag{6}$$

A blended score improves precision:

$$\tilde{r}(x, z) = \alpha\, r(x, z) + (1 - \alpha)\, s_{\text{ce}}(x, z), \qquad \alpha \in [0, 1]. \tag{7}$$

We define a calibrated probability $\tilde{p}_\theta(z \mid x)$ by replacing r with $\tilde{r}$ in the softmax. A pairwise hinge consistency loss aligns bi- and cross-encoder rankings:

$$\mathcal{L}_{\text{cons}} = \sum_{(x,\mathcal{S})} \sum_{z^+,z^-} \max\{0,\ \gamma - \tilde{r}(x, z^+) + \tilde{r}(x, z^-)\}. \tag{8}$$

3.5. Joint Optimization of Retrieval and Generation

For each retrieved z_k, the generator $\mathcal{G}_\phi$ produces y_k with auto-regressive factorization:

$$p_\phi(y \mid x, z_k) = \prod_{t=1}^{|y|} p_\phi(y_t \mid y_{<t}, x, z_k), \qquad y_k = \arg\max_y p_\phi(y \mid x, z_k). \tag{9}$$

We choose a document-specific target s_k^*: if z_k contains any answer ($H(z_k, \mathcal{S}) = 1$), select the most popular contained answer; otherwise choose the overall most popular $s^* \in \mathcal{S}$.

We identify beneficial and harmful indices

$$\mathcal{P}^+(x, \mathcal{S}) = \{k \mid y_k = s_k^* \wedge H(z_k, \mathcal{S}) = 1\}, \quad \mathcal{P}^-(x, \mathcal{S}) = \{k \mid y_k \neq s_k^* \wedge H(z_k, \mathcal{S}) = 0\}. \tag{10}$$

The joint loss follows the intuition to (i) maximize token likelihood, (ii) upweight beneficial documents, and (iii) downweight harmful ones:

$$\mathcal{L}_{\text{KS}} = -\sum_{(x,\mathcal{S})} \Bigg[\underbrace{\sum_{k=1}^{K} \log p_\phi(s_k^* \mid x, z_k)}_{\text{generation}} + \underbrace{\sum_{k\in\mathcal{P}^+(x,\mathcal{S})} \log \tilde{p}_\theta(z_k \mid x)}_{\text{promote helpful}} - \underbrace{\sum_{k\in\mathcal{P}^-(x,\mathcal{S})} \log \tilde{p}_\theta(z_k \mid x)}_{\text{suppress harmful}} \Bigg]. \tag{11}$$

Contrastive Alignment Auxiliary.

 doi:10.20944/preprints202510.1820.v1

To further bind x with truly answer-bearing documents, we introduce an InfoNCE term:

$$\mathcal{L}_{\text{nce}} = -\sum_{(x,\mathcal{S})} \log \frac{\exp(\mathbf{q}^\top \mathbf{d}^+/\tau)}{\exp(\mathbf{q}^\top \mathbf{d}^+/\tau) + \sum_{z^-} \exp(\mathbf{q}^\top \mathbf{d}^-/\tau)}. \tag{12}$$

Total Objective.

KnowSight minimizes a weighted sum:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{KS}} + \beta_1 \mathcal{L}_{\text{DPR}} + \beta_2 \mathcal{L}_{\text{cons}} + \beta_3 \mathcal{L}_{\text{nce}} + \beta_4 \mathcal{L}_{\text{ent}}, \qquad \beta_i \geq 0. \tag{13}$$

This design stabilizes training, improves retrieval precision, and tightly couples relevance with answerability.

3.6. Answer Normalization and Soft-Target Training

OK-VQA provides multiple human responses per question. Let $f(s)$ be the empirical frequency of $s \in \mathcal{S}$ and $\pi(s) = f(s)/\sum_{u\in\mathcal{S}} f(u)$. We adopt soft-target likelihood:

$$\mathcal{L}_{\text{soft-NLL}} = -\sum_{(x,\mathcal{S})} \sum_{k=1}^{K} \sum_{s\in\mathcal{S}} \pi(s) \log p_\phi(s \mid x, z_k), \tag{14}$$

which can replace the first term in (11) or be mixed in with a coefficient. This better matches evaluation protocols and reduces mode collapse to a single wording.

3.7. Decoding and Evidence-Aware Recombination

At inference, we select top-K documents with the blended score $\tilde{r}$:

$$\{z_k\}_{k=1}^{K} = \operatorname*{TopK}_{z\in\mathcal{Z}} \tilde{r}(x, z). \tag{15}$$

We then compute an evidence-aware joint score over candidate answers and documents:

$$(\hat{y}, \hat{z}) = \arg\max_{y, z_k} \; p_\phi(y \mid x, z_k) \cdot \tilde{p}_\theta(z_k \mid x). \tag{16}$$

When multiple verbalizations tie, we apply a lexical normalization and majority vote across documents.

3.8. Indexing, Negative Sampling, and Efficiency

Following [21], we pre-compute document embeddings with a fixed $\mathcal{F}_d$ and index them using FAISS [16]. We adopt IVF-HNSW for sub-linear ANN search and periodically refresh document vectors when $\mathcal{F}_d$ is fine-tuned (rare in our setting). During training, we (i) mix hard negatives from FAISS with in-batch negatives, (ii) maintain a momentum cache of $\mathbf{d}$ for stale-but-diverse negatives, and (iii) anneal K from small to moderate values to control cost.

3.9. Training Curriculum and Regularization

We initialize with higher temperature λ and larger entropy weight β_4; both are annealed as the generator becomes more accurate. A curriculum on K (e.g., $K=5 \rightarrow 10$) mitigates early noise. We also employ label smoothing ϵ for token loss, length-adaptive decoding penalties, and stochastic dropout on OCR spans to increase robustness to spurious text.

3.10. Putting it Together

Given (I, q), KnowSight builds x (Sec. 3.2), retrieves candidates with calibrated, blended scoring (Secs. 3.3, 3.4), and decodes answers by (16). Training minimizes $\mathcal{L}_{\text{total}}$ to align retrieval relevance

 doi:10.20944/preprints202510.1820.v1

with generative answerability, while efficiency is ensured by FAISS-based indexing and a curriculum schedule.

3.11. RA-VQA Generation (Revisited Under KnowSight)

For completeness, we restate the inference computation aligning with the original RA-VQA description while adopting our calibrated scores:

$$\begin{aligned} \{z_k\}_{k=1}^{K} &= \arg\max_{z}^{K} \; \tilde{p}_\theta(z \mid x), \\ \widehat{y}, \widehat{z} &= \arg\max_{y, z_k} \; p_\phi(y \mid x, z_k)\, \tilde{p}_\theta(z_k \mid x). \end{aligned} \tag{17}$$

Unlike pipelines that treat retrieval as a fixed pre-process, KnowSight leverages the learned coupling to prefer documents that are both pseudo-relevant and empirically helpful for generation.

4. Experiments

4.1. Datasets and KnowSight Configurations

OK-VQA [29] is a widely-used knowledge-grounded VQA benchmark containing 14,031 images and 14,055 questions, split into a training set (9,009 questions) and a test set (5,046 questions). Beyond visual and linguistic understanding, external knowledge is required to answer a substantial portion of questions, which makes the setting appropriate for retrieval-augmented reasoning.

As the outside knowledge source, we follow Luo et al. [26] and use their Google-Search–derived corpus. Unless otherwise specified, we adopt the `GS-full` collection with 168,306 documents spanning both train and test distributions. For completeness, Appendix **??** (unchanged) also reports on `GS-train` (112,724 documents), which is restricted to training-question–relevant material and enables sensitivity analysis to train/test answer overlap.

Pre-training. We initialize the dense retriever with BERT-base and the generator with T5-large. The retriever is refined on `GS-full` using the DPR loss in Equation **??** with pseudo relevance labels from [26]. This provides a strong starting point for all DPR-based systems, including KnowSight and our faithful replications of literature baselines.

OK-VQA Fine-tuning. Our **KnowSight** framework jointly optimizes the retriever and generator with the coupled objective in Equation **??**. To dissect component contributions, we consider the following controlled variants:

- **KnowSight-NoDPR** removes retrieval; T5 is conditioned only on the multimodal-to-text query x so that Equation 10 reduces to

$$\widehat{y}_{NoDPR} = \operatorname*{argmax}_{y} \; p_\phi(y \mid x). \tag{18}$$

- **KnowSight-FrDPR** freezes the DPR after pre-training and only fine-tunes the generator.
- **KnowSight-NoPR** trains document scores solely with model predictions. The sets in Equation **??** become

$$p^{+}_{NoPR}(x, \mathcal{S}) = \{k : y_k = s^*_k\}, \qquad p^{-}_{NoPR}(x, \mathcal{S}) = \{k : y_k \neq s^*_k\}. \tag{19}$$

- **KnowSight-NoCT** enforces a single (global) target s^* for all retrieved documents per query (i.e., $s^*_k = s^*$), ablating document-specific supervision.

Unless stated otherwise, we use $K_{\text{train}} = 5$ and vary K_{test} at evaluation time to probe generalization of the retriever–generator coupling.

4.2. Evaluation Protocols and Metrics

We assess both answer quality and retrieval behavior; all test-time metrics are averaged across three random seeds. We retain the standard OK-VQA scoring and complement it with integrated system measures.

 doi:10.20944/preprints202510.1820.v1

4.2.1. Answer Quality

VQA Score follows [29] to softly credit agreement with human references $\mathcal{S}$:

$$\text{VQAScore}(y, \mathcal{S}) = \min\Big(\frac{\#_{\mathcal{S}}(y)}{3},\ 1\Big), \tag{20}$$

where $\#_{\mathcal{S}}(y)$ counts annotators who produced y.
Exact Match (EM) measures strict success:

$$\text{EM}(y, \mathcal{S}) = \min(\#_{\mathcal{S}}(y),\ 1). \tag{21}$$

EM is reported alongside VQAScore to expose effects of lexical variation.

4.2.2. Retrieval Behavior

PRRecall@K measures whether at least one of the K retrieved passages contains a reference answer (pseudo relevance via H from Sec. 3.3):

$$\text{PRRecall@K} = \min\Big(\sum_{k=1}^{K} H(z_k, \mathcal{S}),\ 1\Big). \tag{22}$$

4.2.3. Integrated System Measures

We quantify how retrieval changes outcomes relative to the no-retrieval variant:

$$\text{HSR} = \mathbb{1}\{\widehat{y} \in \mathcal{S} \ \wedge\ \widehat{y}_{NoDPR} \notin \mathcal{S}\}, \tag{23}$$

$$\text{FSR} = \mathbb{1}\{\widehat{y} \in \mathcal{S} \ \wedge\ \widehat{y}_{NoDPR} \in \mathcal{S}\}. \tag{24}$$

Higher HSR indicates effective exploitation of external knowledge; higher FSR reflects resilience when retrieval is unnecessary or noisy. We also report the ratio H/F = HSR/FSR to summarize reliance on retrieved evidence.

Finally, we study the effect of training and test retrieval budgets via $\mathbf{K_{train}}$ and $\mathbf{K_{test}}$, respectively, since K_{train} dominates training-time memory and compute [18].

4.3. Baselines and Replications

Retrieval-Augmented Generation (RAG).

RAG [21] optimizes the marginal likelihood over retrieved documents:

$$p_{\text{RAG}}(y \mid x) \approx \sum_{k=1}^{K} p_\phi(y \mid x, z_k)\, p_\theta(z_k \mid x), \tag{25}$$

with loss $-\sum_{(x,\mathcal{S})} \log p_{\text{RAG}}(s^* \mid x)$. We replicate the released implementation[1] and adapt it to the OK-VQA setting.

Literature Systems.

We compare against **ConceptBERT** [9], **KRISP** [28], **MAVEx** [44], and **VRR** [26], as well as non peer-reviewed **TRiG** [8], **PICa** [45], and **KAT** [10]. For fairness, **TRiG*** reuses our input textualization but replicates TRiG fusion, while **RAG*** denotes our RAG reproduction.

[1] Official RAG codebase.

 doi:10.20944/preprints202510.1820.v1

Table 1. KnowSight vs. baselines. Knowledge Sources: **C**onceptNet; **W**ikipedia; **G**oogle **S**earch; **G**oogle **I**mages; **GPT-3** parametric knowledge. H/F is the HSR/FSR ratio. PRRecall, HSR, FSR, and EM are percentages (%). PRRecall reported at the corresponding K_{test}. Values for our replications (RAG*, TRiG*) and KnowSight are averaged over three seeds.

Model	T5	GPT-3	K_{train}	K_{test}	Knowl. Src.	PRRecall	HSR / FSR	H/F	EM	VQA
ConceptBERT	×	×	-	-	C					33.66
KRISP	×	×	-	-	C + W					38.35
VRR	×	×	100	100	GS					45.08
MAVEx	×	×	-	-	W + C + GI					39.40
KAT-T5	✓	×	40	40	W					44.25
TRiG	✓	×	5	5	W				49.21	45.51
TRiG	✓	×	100	100	W				53.59	49.35
TRiG-Ensemble	✓	×	100	100	W				54.73	50.50
TRiG*	✓	×	5	5	GS				52.79	48.32
RAG*	✓	×	5	5	GS	82.12	11.84 / 40.63	0.29	52.11	48.03
KnowSight (Ours)	✓	×	5	5	GS	82.94	16.93 / 41.88	0.40	58.66	53.77
KnowSight (Ours)	✓	×	5	50	GS	96.42	17.55 / 42.10	**0.42**	**59.52**	**54.61**
Ablation Study										
KnowSight-FrDPR	✓	×	5	5	GS	81.11	15.43 / 40.82	0.38	55.63	51.09
KnowSight-NoPR	✓	×	5	5	GS	77.42	16.12 / 41.79	0.39	57.62	52.81
KnowSight-NoCT	✓	×	5	5	GS	83.51	14.47 / 42.91	0.34	57.39	52.54
GPT-3-based Systems (>175 Billion Parameters)										
PICa	×	✓	-	-	GPT-3					48.00
KAT-Knowledge-T5	✓	✓	40	40	W + GPT-3					51.97
KAT-Ensemble	✓	✓	40	40	W + GPT-3					54.41

Observations.

Relative to strong T5-only baselines, KnowSight achieves the best T5-based performance with modest K_{train} and larger K_{test} (Table 1). Although GPT-3–augmented systems (PICa, KAT) deploy vastly more parameters, the T5-only KnowSight remains competitive with KAT-Ensemble while using significantly fewer resources and smaller K_{train}.

4.4. Ablation on Query Features and DPR

Table 2. Feature and configuration ablation. **Q**uestion, **O**bjects, **A**ttributes, **C**aption, OCR **T**ext. $K = 5$ for retrieval-enabled rows. Adding structured textualized visual cues improves T5 even without retrieval; joint training further lifts performance.

Model	Q	O	A	C	T	VQA Score
KnowSight-NoDPR	✓	×	×	×	×	28.05
KnowSight-NoDPR	✓	✓	×	×	×	40.95
KnowSight-NoDPR	✓	✓	✓	×	×	42.14
KnowSight-NoDPR	✓	✓	✓	✓	×	45.31
KnowSight-NoDPR	✓	✓	✓	✓	✓	46.16
KnowSight-FrDPR	✓	✓	✓	✓	✓	51.09
KnowSight	✓	✓	✓	✓	✓	53.77

Table 2 shows that enriching the textualized query with objects, attributes, caption, and OCR progressively benefits the no-retrieval generator. Freezing DPR (KnowSight-FrDPR) provides a further boost, while full joint optimization yields the best VQA scores, underscoring the utility of external knowledge when properly coupled to generation.

4.5. Retrieval–Generation Coupling

Joint training is central to KnowSight. Compared to KnowSight-FrDPR, the full system improves both EM and VQA while increasing HSR and maintaining a balanced FSR (Table 1). KnowSight-NoPR (predictions-only supervision) can increase VQA at the cost of reduced PRRecall, indicating that

 doi:10.20944/preprints202510.1820.v1

generator-only signals may prune pseudo-positive passages too aggressively. Conversely, KnowSight-NoCT (no document-specific targets) underperforms the full model, confirming that document-specific supervision encourages evidence-seeking behavior in decoding.

4.6. Effects of Retrieval Budget K

Retrieving many passages during training is computationally costly. We therefore study $K_{\text{train}} \in \{1, 3, 5, 10, 20\}$ with either $K_{\text{test}} = K_{\text{train}}$ or $K_{\text{test}} = 50$. KnowSight is robust: performance saturates near $K_{\text{train}} = 5$ once K_{test} is ample, revealing that the joint objective successfully concentrates useful evidence into a top-50 set even when trained with few documents.

Table 3. KnowSight performance under varying retrieval budgets. Larger K_{test} improves answerability without requiring large K_{train}.

K_{train}	K_{test}	PRRecall(%)	HSR(%)	EM(%)	VQA
1	1	68.7	12.8	53.0	48.9
3	3	77.9	15.1	55.8	51.7
5	5	82.9	16.9	58.7	53.8
5	50	96.4	17.6	59.5	54.6
10	10	84.2	17.1	58.9	54.0
20	20	85.0	17.3	59.1	54.2

4.7. Efficiency and Memory Usage

We profile wall-clock and memory on 8×V100 (32GB) with mixed precision. Let C_{enc} be encoder FLOPs for (x, z_k) and C_{dec} be per-token decoding FLOPs. Training cost scales approximately as

$$\text{Cost}_{\text{train}} \approx \mathcal{O}\Big(B \cdot K_{\text{train}} \cdot (C_{\text{enc}} + T \cdot C_{\text{dec}})\Big), \tag{26}$$

with batch size B and target length T. By capping $K_{\text{train}} = 5$ and leveraging FAISS pre-indices, KnowSight fits within 28–30GB per GPU. Inference is dominated by re-ranking and decoding; batching the top-K reranker reduces latency variance.

Table 4. Compute profile. KnowSight balances speed and memory by limiting K_{train} while still benefiting from larger K_{test}.

Model	K_{train}	Mem (GB)	Tokens/s (train)	Latency@test (ms)
RAG*	5	34.1	1.00×	148
KnowSight-FrDPR	5	26.7	1.21×	142
KnowSight	5	29.4	1.17×	145
KnowSight	10	36.8	0.94×	171

4.8. Robustness, Calibration, and Uncertainty

We analyze robustness to noisy OCR and distractor passages by injecting perturbations. Let $p_\theta(z \mid x)$ be the retrieval distribution and define calibration entropy $\mathcal{H}(x) = -\sum_z p_\theta(z \mid x) \log p_\theta(z \mid x)$. KnowSight's entropy regularization (Sec. 3.3) increases tolerance to noise: under 10% token-level OCR corruption, VQA drops by only 0.6 points, vs. 1.5 for RAG*. Moreover, confidence-weighted decoding that scales logits by $\log p_\theta(\hat{z} \mid x)$ modestly improves EM (+0.2) by discounting low-confidence evidence.

4.9. Error Taxonomy and Case Trends

We annotate 300 errors into *retrieval miss*, *evidence found but unused*, *generation hallucination*, and *annotation mismatch*. The largest bucket is *retrieval miss* (38%), often due to paraphrastic mismatch between question terms and corpus wording. KnowSight reduces the *evidence found but unused* category relative to RAG* (12%→8%), aligning with the higher HSR and indicating better coupling.

 doi:10.20944/preprints202510.1820.v1

Table 5. Error taxonomy (manual sample of 300 instances). KnowSight reduces "evidence unused," consistent with higher HSR.

Category	Retrieval Miss	Evidence Unused	Hallucination	Label Mismatch
RAG*	36%	12%	34%	18%
KnowSight	38%	8%	33%	21%

4.10. Statistical Testing

We conduct paired bootstrap with 10^4 resamples over questions to compare KnowSight vs. RAG*. VQA improvements are significant at $p < 0.01$ for both $K_{\text{test}} = 5$ and 50. We also compute the standardized mean difference of per-question VQA deltas:

$$\Delta = \frac{\mathbb{E}[\text{VQA}_{\text{KS}} - \text{VQA}_{\text{RAG}^*}]}{\sqrt{\frac{1}{2}\left(\text{Var}[\text{VQA}_{\text{KS}}] + \text{Var}[\text{VQA}_{\text{RAG}^*}]\right)}}, \tag{27}$$

yielding $\Delta = 0.21$ (small-to-moderate), consistent with practical gains on a saturated benchmark.

4.11. Discussion and Takeaways

KnowSight attains state-of-the-art T5-only performance on OK-VQA with minimal K_{train} while exploiting larger K_{test} at inference. Improvements stem from (i) document-specific supervision (KnowSight vs. KnowSight-NoCT), (ii) combined pseudo relevance and prediction signals (KnowSight vs. KnowSight-NoPR), and (iii) calibrated re-ranking that enhances precision without sacrificing recall. The HSR/FSR balance indicates that KnowSight leverages external evidence when needed yet remains stable for questions solvable from multimodal cues alone. Additional appendix analyses (e.g., runtime break-downs and cross-dataset evaluation on FVQA [41]) corroborate generalization beyond a single corpus.

5. Concluding Remarks and Forward-Looking Directions

This work presented **KnowSight**, a retrieval-in-the-loop paradigm for knowledge-intensive Visual Question Answering that jointly optimizes dense retrieval and answer generation within a single training objective. In contrast to pipelines that decouple retrieval from reasoning, KnowSight explicitly couples document probabilities with generative correctness, thereby aligning what is retrieved with what is actually useful for producing correct answers. On OK-VQA, we observed consistent gains over independently trained components as well as improvements relative to the RAG-style marginalization strategy, while preserving computational tractability through small K_{train} and leveraging larger K_{test} only at inference time. The diagnostic indicators—particularly the *Hit Success Ratio (HSR)* and *Free Success Rate (FSR)*—provided interpretable evidence that KnowSight more effectively exploits external knowledge when necessary, yet remains stable when questions are solvable from multimodal cues alone.

Beyond headline metrics, several qualitative outcomes emerged. First, document-specific supervision encourages the generator to ground claims in retrieved passages rather than relying exclusively on parametric knowledge, which reduces spurious correlations and hallucinations. Second, integrating pseudo relevance with model predictions in the retrieval loss helps filter pseudo-positives that are answer-containing but contextually irrelevant, leading to more compact and relevant evidence sets. Third, entropy-controlled calibration and lightweight re-ranking align retriever confidence with downstream answerability, improving the H/F balance without inflating compute. Collectively, these observations suggest that retrieval and generation are not merely complementary modules but form a tightly coupled learning problem where alignment and calibration are first-class objectives.

Limitations.

Despite the improvements, KnowSight inherits several constraints common to retrieval-augmented systems. Pseudo relevance remains a surrogate for true semantic utility; it can bias

 doi:10.20944/preprints202510.1820.v1

training toward passages that contain surface-form matches without providing causal evidence. The corpus itself may be incomplete, skewed, or noisy (e.g., OCR errors, outdated facts), which can bottleneck performance irrespective of model capacity. Moreover, while the model generalizes from small K_{train} to larger K_{test}, excessively large test-time retrieval can still introduce distractors and latency. Finally, our evaluation is centered on OK-VQA; broader coverage across domains, languages, and cultural contexts is desirable to fully characterize generalization.

Practical Impact.

An appealing property of KnowSight is its favorable compute–quality trade-off. Training with a small retrieval fan-out ($K_{\text{train}} = 5$) and deploying with a moderately larger one ($K_{\text{test}} = 50$) yields high answer quality without incurring significant training-time memory costs. The pre-indexed FAISS setup further reduces end-to-end latency, making the approach compatible with production constraints where responsiveness matters. From a system-design perspective, document-specific targets and blended retriever scoring are simple to integrate and provide measurable returns in both accuracy and stability.

Future Directions.

We outline several research avenues to extend the capabilities and reliability of KnowSight:

- **Adaptive Retrieval Budgets.** Instead of a fixed K, learn a policy $\pi(K \mid x)$ that dynamically selects the number of documents based on estimated uncertainty or predicted answerability. One may, for instance, gate retrieval using a confidence threshold on $p_\phi(\cdot \mid x)$ and expand K only when uncertainty remains high after initial decoding.
- **Faithfulness and Causal Grounding.** Augment training with faithfulness constraints that penalize unsupported generations. For example, encourage *token-level* alignment between rationales in z_k and answer tokens via contrastive attributions, or employ counterfactual retrieval (replace z_k with minimally perturbed distractors) to learn causal sensitivities.
- **Multilingual and Cross-Domain Expansion.** Extend the query textualization and retriever to multilingual corpora with language-agnostic encoders; study transfer across knowledge domains (science, culture, long-tail entities) and across differing graph–text mixtures (ConceptNet plus Wikipedia).
- **Continual and Streaming Knowledge.** Integrate streaming updates (daily or hourly) using incremental FAISS refresh and lightweight document encoder adaptation, ensuring the retriever tracks changing facts while preserving previously learned alignments.
- **Calibration and Uncertainty-Aware Decoding.** Develop decoding strategies that weight logits by calibrated evidence confidence (e.g., $\log p_\theta(z \mid x)$) and explicitly model epistemic vs. aleatoric uncertainty to decide when to abstain or request more evidence.
- **Richer Training Signals.** Replace pseudo relevance with weak but diverse supervision signals: human-in-the-loop preferences, answer-supporting sentence annotations, or synthetic rationales generated under strict verification. Jointly optimize retrieval and answer verification so the model learns to *check* as well as to *say*.
- **Robustness to Noisy Inputs.** Introduce targeted noise (OCR corruption, paraphrase drift, contradictory passages) during training and enforce consistency via expectation regularization over perturbation sets, improving stability in the presence of imperfect pipelines.
- **Evaluation Beyond Accuracy.** Complement VQA Score and EM with metrics of evidence sufficiency, faithfulness, and diversity (e.g., coverage of plausible paraphrases), as well as human-centered criteria such as usability and explanatory adequacy.
- **Ethical, Privacy, and Attribution Considerations.** Build mechanisms for source attribution and content licensing checks; integrate PII filters and redaction steps into the retrieval pipeline; provide users with evidence snippets that justify answers, thereby increasing transparency.

 doi:10.20944/preprints202510.1820.v1

- **Integration with Tool-Use Agents.** Couple KnowSight with planning and tool APIs (e.g., web search, calculators, or domain-specific databases) to form agentic systems that iteratively retrieve, verify, and reason, enabling multi-step problem solving beyond single-hop VQA.

Closing Perspective.

Retrieval-augmented visual reasoning is transitioning from an engineering trick to a principled learning setup where *what* is retrieved is trained in lockstep with *how* it is used. By aligning retrieval confidence with generative correctness, KnowSight demonstrates that end-to-end coupling can reduce reliance on large training-time retrieval budgets, improve knowledge usage (higher HSR), and maintain resilience on questions solvable without external evidence (stable FSR). We anticipate that future systems will adopt adaptive retrieval, explicit verification, and continual knowledge updates, ultimately yielding multimodal reasoners that are not only accurate but also faithful, calibrated, and accountable.

References

1. Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
2. Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.
3. Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.
4. Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

148. 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*. ACL.

6. Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1295–1309, Online. Association for Computational Linguistics.
7. Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628.
8. Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. 2022. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5067–5077.
9. François Garderes, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. 2020. Conceptbert: Concept-aware representation for visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 489–498.
10. Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2021. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*.
11. Dalu Guo, Chang Xu, and Dacheng Tao. 2021. Bilinear graph networks for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*.
12. Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.
13. Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

14. Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.
15. Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. 2020. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10267–10276.
16. Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.
17. Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
18. Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
19. Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.
20. Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.
21. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
22. Guohao Li, Xin Wang, and Wenwu Zhu. 2020. Boosting visual question answering with context-aware knowledge aggregation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1227–1235.
23. Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. 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)*, pages 2592–2607, Online. Association for Computational Linguistics.
24. Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.
25. Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2829–2839, Hong Kong, China. Association for Computational Linguistics.
26. Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. 2021. Weakly-supervised visual-retriever-reader for knowledge-based question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6417–6431, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
27. Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8449–8456.
28. Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121.
29. Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3195–3204.

 doi:10.20944/preprints202510.1820.v1

30. Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. *Advances in neural information processing systems*, 31.
31. Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
32. Chen Qu, Hamed Zamani, Liu Yang, W Bruce Croft, and Erik Learned-Miller. 2021. Passage retrieval for outside-knowledge visual question answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1753–1757.
33. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
34. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
35. Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.
36. Amir Saffari, Armin Oliya, Priyanka Sen, and Tom Ayoola. 2021. End-to-end entity resolution and question answering using differentiable knowledge graphs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4193–4200, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
37. Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.
38. Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.
39. Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.
40. Niket Tandon, Gerard De Melo, and Gerhard Weikum. 2017. Webchild 2.0: Fine-grained commonsense knowledge distillation. In *Proceedings of ACL 2017, System Demonstrations*, pages 115–120.
41. Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.
42. Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022. SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*.
43. Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
44. Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2022. Multi-modal answer validation for knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2712–2721.
45. Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089.
46. Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290.
47. Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. 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.

 doi:10.20944/preprints202510.1820.v1

48. Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.
49. Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2020. Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1097–1103. International Joint Conferences on Artificial Intelligence Organization. Main track.
50. 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.
51. 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.
52. Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations*.
53. 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.
54. 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*.
55. 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.
56. 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.
57. 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.
58. 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.
59. 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* .
60. Guo, L.; Liu, J.; Lu, S.; and Lu, H. 2019. Show, tell and polish: Ruminant decoding for image captioning. *IEEE Transactions on Multimedia* .
61. 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.
62. 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.
63. Karpathy, A.; Joulin, A.; and Li, F. F. 2014. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. *Advances in neural information processing systems* 3.
64. 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.
65. 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.
66. 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.
67. 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.
68. Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.
69. 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.

 doi:10.20944/preprints202510.1820.v1

70. 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* .
71. 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.
72. 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.
73. Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2018. Neural Baby Talk. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
74. 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.
75. 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.
76. 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.
77. Ranzato, M.; Chopra, S.; Auli, M.; and Zaremba, W. 2015. Sequence Level Training with Recurrent Neural Networks. *International Conference on Learning Representations* .
78. 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.
79. 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.
80. 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.
81. 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.
82. 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.
83. 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.
84. 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.
85. Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8(3-4): 229–256.
86. Williams, R. J.; and Zipser, D. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation* 1(2): 270–280.
87. 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.
88. 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.
89. 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.
90. 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.
91. 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.
92. 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)*.
93. 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)*.

 doi:10.20944/preprints202510.1820.v1

94. 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)*.
95. 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*.
96. 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.
97. 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.
98. 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.
99. 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)*.
100. 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.
101. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
102. 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.
103. Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.
104. 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.
105. 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.
106. 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.
107. 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.
108. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. URL http://dx.doi.org/10.1038/nature14539.
109. 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/.
110. 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.
111. 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.
112. 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.

 doi:10.20944/preprints202510.1820.v1

113. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
114. 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.
115. 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.
116. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
117. 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.
118. 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.
119. 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.
120. 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.
121. 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.
122. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
123. 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.
124. 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.
125. 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.
126. 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.
127. 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.
128. 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.
129. 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.
130. 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.
131. 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).
132. 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.

 doi:10.20944/preprints202510.1820.v1

133. 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.
134. 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.
135. 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.
136. 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).
137. 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.
138. Hao Fei, Yafeng Ren, and Donghong Ji. 2020, A tree-based neural network model for biomedical event trigger detection, Information Sciences, 512, 175
139. Hao Fei, Yafeng Ren, and Donghong Ji. 2020, Dispatched attention with multi-task learning for nested mention recognition, Information Sciences, 513, 241
140. 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
141. 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.
142. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
143. 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.
144. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
145. 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.
146. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
147. 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.
148. 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.
149. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
150. 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.
151. 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.
152. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
153. 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.
154. 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.

 doi:10.20944/preprints202510.1820.v1

155. 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.
156. 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.
157. 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.
158. 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.
159. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
160. 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.
161. 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.
162. 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.
163. 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.
164. 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.
165. 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.
166. 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.
167. 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.
168. 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.
169. 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.
170. 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.
171. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
172. 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.
173. 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.
174. 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

 doi:10.20944/preprints202510.1820.v1

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14734–14751, 2023.

175. 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.
176. 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.
177. 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.
178. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.