KR-VLM: Enhancing Factual Reasoning in Vision-Language Models via Knowledge Retrieval and Self-Verification

Jie-Hao Lim, Carter Ross
Singapore Institute of Management

Abstract

Vision-Language Models (VLMs) exhibit powerful capabilities in visual and textual understanding, significantly advancing tasks like Visual Question Answering (VQA). However, hallucination remains a persistent challenge, as VLMs generate responses factually inconsistent with the input image or common sense. This undermines their reliability and trustworthiness, especially in scenarios demanding precise factual reasoning or complex scene comprehension. To address this, we propose KR-VLM (Knowledge-Retrieved Reasoning for Vision-Language Models), a novel approach enhancing VLM factual reasoning and significantly reducing hallucination via knowledge-aware self-supervised learning. KR-VLM integrates a Knowledge Retrieval Module (KRM) to access external facts, a Knowledge Fusion & Calibration Adapter (KFCA) to seamlessly integrate cross-modal knowledge, and a Self-Factual Verification Module (SFVM) to self-correct factual inconsistencies during training. Leveraging an established VLM architecture, our method is lightweight and requires no extensive human annotation of knowledge or reasoning paths. Extensive experiments on VQAv2, GQA, OK-VQA, and DocVQA benchmarks show that KR-VLM consistently outperforms state-of-the-art baselines in VQA accuracy and, crucially, achieves a superior Factual Consistency Score, demonstrating its effectiveness in mitigating hallucination.

1 Introduction

Vision-Language Models (VLMs) have emerged as a cornerstone of modern artificial intelligence, bridging the gap between perception and language understanding [1, 2, 3]. Their ability to process and reason over multimodal inputs has led to groundbreaking performance in various tasks, including image captioning, visual dialogue, and particularly Visual Question Answering (VQA) [4, 5]. These models have demonstrated impressive generalization capabilities, often exhibiting emergent behaviors that hint at a deeper understanding of the world [6]. The significance of VLMs lies in their potential to power intelligent systems that can interact with humans more naturally, assist in complex analytical tasks, and augment decision-making processes across diverse domains, from healthcare [7, 8, 9], fraud detection [10] to autonomous driving [11, 12, 13].

Despite these significant strides, current VLMs are plagued by a critical limitation: *hallucination*. Hallucination refers to the phenomenon where models generate information that is plausible but factually incorrect or inconsistent with the visual evidence presented in the input image. For instance, a VLM might claim there are two red apples when the image clearly shows three green ones, or state that a person is wearing a hat when they are not. This issue is particularly pronounced in tasks demanding precise factual recall or intricate logical reasoning, where even minor inaccuracies can render the model's output unreliable and untrustworthy. Such factual inaccuracies severely impede the deployment of VLMs in high-stakes applications where correctness is paramount, necessitating robust solutions to enhance their factual grounding, especially when dealing with chaotic or complex contexts [14].

Addressing VLM hallucination presents several existing challenges. While Large Language Models (LLMs) have successfully leveraged Retrieval-Augmented Generation (RAG) techniques to improve factual accuracy by consulting external knowledge bases, seamlessly integrating such external knowledge

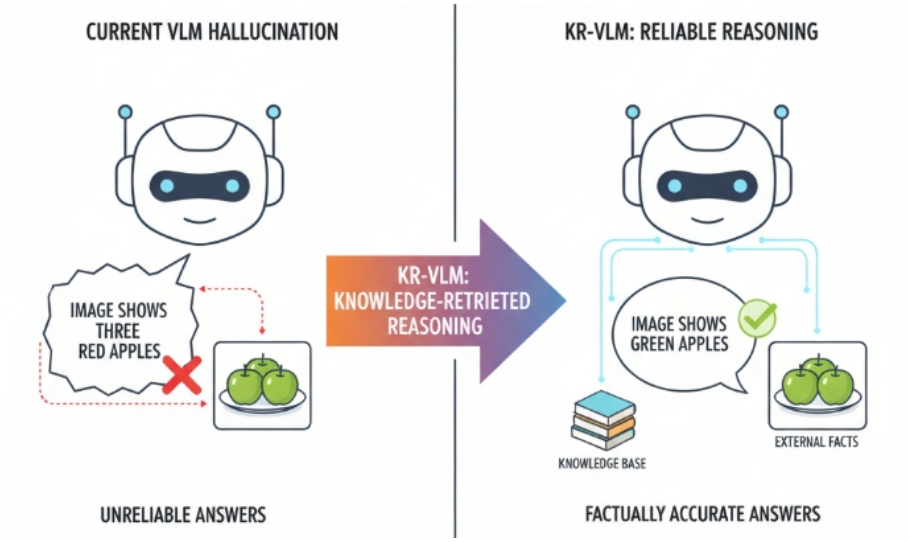

Figure 1: Addressing VLM Hallucination: KR-VLM leverages knowledge retrieval and reasoning to transform unreliable answers into factually accurate responses.

into the cross-modal reasoning pipeline of VLMs remains an open problem. Directly adapting RAG to VLMs often entails prohibitively high re-training costs, significant inference latency, or an inability to effectively fuse knowledge across modalities. Traditional fine-tuning approaches, while improving performance on specific datasets, typically fail to fundamentally eliminate hallucination and often exhibit poor adaptability to domain-specific knowledge. Our motivation stems from the need for a lightweight and efficient method that can intrinsically enhance a VLM's factual reasoning capacity by harnessing external knowledge without requiring extensive architectural modifications or expensive human annotations.

To this end, we propose **KR-VLM (Knowledge-Retrieved Reasoning for Vision-Language Models)**, a novel framework designed to mitigate hallucination and bolster factual consistency in VLMs. Built upon the robust LLaVA-1.6 architecture (Vision Encoder: CLIP-ViT-L-336px, Language Decoder: Mistral 7B), KR-VLM introduces three key components: **Knowledge Retrieval Module (KRM)** that, in parallel to visual processing, queries a pre-built external knowledge base (e.g., Wikipedia, ConceptNet) to retrieve factual information relevant to the image content and the user's question. This module employs a dual-encoder architecture for efficient similarity search, ensuring that critical facts are identified early in the reasoning process. **Knowledge Fusion & Calibration Adapter (KFCA)**, a lightweight, cross-modal adapter strategically placed within the VLM's language decoder layers. The KFCA is responsible for dynamically integrating the retrieved knowledge representations (e.g., text embeddings) with the VLM's internal visual and linguistic representations. It utilizes gating or attention mechanisms to carefully modulate the influence of external knowledge, preventing over-reliance while ensuring factual accuracy. **Self-Factual Verification Module (SFVM)**, which operates during the training phase as an auxiliary component. The SFVM evaluates the factual consistency of the VLM's intermediate reasoning steps and final generated answers against both the retrieved knowledge and known ground-truth facts. It provides a "soft label" form of factual feedback by checking for entailment or contradiction, acting as a regularization signal to guide the KFCA's learning and ensure generated outputs are factually aligned. This integrated approach allows KR-VLM to self-calibrate and reinforce its understanding and generation of factual information by explicitly leveraging and verifying external knowledge, without significantly increasing model parameters or inference time.

Our experimental evaluation is conducted on a suite of widely recognized VQA benchmarks known for their demands on factual reasoning: VQAv2, GQA, OK-VQA, and DocVQA. We also include InfographicVQA to test performance on specialized document understanding. A notable advantage of our method is its reliance on knowledge-aware self-supervised learning, which means it *does not* require new

human annotations for reasoning steps or external knowledge. Instead, samples are drawn directly from existing VQA datasets, and external knowledge bases are constructed independently, with the SFVM providing self-generated "soft supervision."

We evaluate KR-VLM's performance using standard VQA metrics such as Accuracy and Normalized Edit Distance Similarity (ANLS), alongside a newly introduced **Factual Consistency Score (FCS)**. The FCS quantifies the reduction in hallucination by measuring the factual alignment between generated answers and reference answers. Our results demonstrate that KR-VLM consistently achieves superior performance across all evaluation metrics compared to baseline methods, including the vanilla LLaVA-1.6 and other knowledge-enhanced approaches. Specifically, KR-VLM shows a notable improvement in the Factual Consistency Score, confirming its efficacy in suppressing hallucination. For instance, KR-VLM achieves an FCS of 76.1%, significantly outperforming LLaVA-1.6 (68.3%) and a baseline with post-hoc knowledge checking (72.8%). This highlights the profound impact of our integrated knowledge-retrieval, fusion, and self-verification mechanism. Furthermore, ablation studies confirm the crucial contribution of each proposed module, particularly the KRM and SFVM, to the overall performance gains in factual accuracy.

In summary, our key contributions are:

- We propose KR-VLM, a novel framework that effectively integrates external factual knowledge and self-supervision to significantly reduce hallucination and enhance the factual consistency of VLM outputs.
- We introduce a lightweight and efficient architecture comprising a Knowledge Retrieval Module, a Knowledge Fusion & Calibration Adapter, and a Self-Factual Verification Module, designed for seamless integration into existing VLM pipelines without extensive parameter overhead.
- We demonstrate that KR-VLM achieves state-of-the-art performance on multiple VQA benchmarks, notably improving factual accuracy and consistency while obviating the need for manual knowledge annotations, thus promoting scalable and reliable VLM development.

2 Related Work

2.1 Vision-Language Models and Hallucination Mitigation

Efforts to enhance Vision-Language Models (VLMs) and mitigate their propensity for hallucination span various fronts. Huang et al. [15], for instance, tackled performance degradation in zero-shot cross-lingual VLM transfer by proposing a multilingual multimodal pre-training strategy and introducing the MultiHowTo100M dataset, which significantly improved cross-lingual text-to-video search capabilities. To directly address hallucination through improved grounding, ViGoRL [16] presents a reinforcement learning-trained VLM that explicitly anchors each reasoning step to specific visual coordinates, thereby achieving spatially grounded multimodal learning. This approach enhances visual verification and accurate localization, dynamically guiding attention to relevant visual regions. Evaluating such advanced reasoning is crucial, as exemplified by GeoQA [17], a benchmark specifically designed to assess multimodal numerical and geometric reasoning capabilities within VQA systems, which is critical for addressing complex hallucinations related to quantitative understanding. Beyond specific VLM architectures, Rawte et al. [18] offer a comprehensive framework for understanding and mitigating hallucination in Large Language Models, detailing a fine-grained taxonomy and introducing tools like a hallucination elicitation dataset and a vulnerability index, providing valuable insights applicable across generative AI, including VLMs. Recent advancements in general computer vision tasks, such as dynamic memory for video object segmentation [19], open-vocabulary segmentation with semantic calibration [20], and universal segmentation guided by language instructions [21], further underscore the growing capabilities and complexity of vision models, some of which leverage language for enhanced performance. These capabilities extend to robust perception for robotics [22] and efficient low-light video processing [23]. Furthermore, Song et al. [24] demonstrated CLIP's efficacy as a few-shot VLM for generative VQA tasks, illustrating how parameter-efficient learning in foundational VLM encoders can foster data-efficient and

robust generative models, indirectly aiding hallucination mitigation. Enhancing VLM reliability further, the Cross-Modal Associative Learning (CMAL) framework [25] achieves robust cross-modal alignment with significantly fewer training corpora compared to traditional contrastive methods, promoting a more resource-efficient and reliable learning process less susceptible to extensive data dependencies. Finally, the broader challenges of trustworthiness are highlighted by Meade et al. [26], who surveyed bias mitigation techniques in pre-trained language models and found inherent trade-offs with core model capabilities, underscoring the complexities in enhancing overall trustworthiness and mitigating undesirable behaviors like hallucinations in VLMs.

2.2 Retrieval-Augmented and Knowledge-Aware Models for Factual Reasoning

The development of Retrieval-Augmented and Knowledge-Aware Models is crucial for enhancing factual reasoning and addressing limitations in parametric knowledge. Mao et al. [27] tackle challenges in multi-hop factual reasoning within Retrieval-Augmented Generation (RAG), particularly regarding irrelevant retrieved passages and error propagation, by introducing Tree of Reviews (ToR), a dynamic tree-based retrieval framework designed to independently evaluate and manage retrieved paragraphs, thus enhancing retrieval robustness and response generation accuracy. To further unify knowledge integration, Oguz et al. [28] presented UniK-QA, a framework that unifies representations of structured and unstructured knowledge, proving highly relevant for knowledge-aware models that integrate external knowledge bases for factual reasoning through techniques such as large pre-trained language models and effective data sampling. Similarly, Yasunaga et al. [29] introduced QA-GNN, a model that explicitly integrates external knowledge from knowledge graphs with language models to enhance factual reasoning in question answering. Addressing the limitations of static knowledge bases, particularly in retrieval-augmented code generation, the EVOR pipeline [30] introduces synchronously evolving queries and diverse knowledge bases, significantly improving execution accuracy and enhancing models' ability to retrieve and apply correct factual information for complex programming tasks. The necessity of such external knowledge sources is underscored by Mallen et al. [31], who demonstrated that large language models struggle with less popular factual knowledge where parametric memory is insufficient, highlighting the critical role of integrating non-parametric knowledge, like Knowledge Graphs, for comprehensive factual reasoning, especially for long-tail information. Benchmarks like TemporalWiki [32] are highly relevant for advancing these models by enabling the development of self-supervised learning methods for continuous knowledge adaptation in ever-evolving language models. Furthermore, to enhance domain-agnostic factual reasoning and mitigate hallucination, Longpre et al. [33] and Chen et al. [34] demonstrated the effectiveness of data sampling strategies, particularly negative sampling. This technique trains models to discern unanswerable questions, thereby preventing the generation of unfounded responses and improving performance on factual verification tasks.

3 Method

We introduce **KR-VLM (Knowledge-Retrieved Reasoning for Vision-Language Models)**, a novel framework designed to enhance the factual reasoning capabilities of Vision-Language Models (VLMs) and significantly mitigate the issue of hallucination. Built upon the robust LLaVA-1.6 architecture, KR-VLM strategically integrates external knowledge into the VLM's cross-modal reasoning pipeline through a lightweight and self-supervised approach. Our method comprises three core modules: the Knowledge Retrieval Module (KRM), the Knowledge Fusion & Calibration Adapter (KFCA), and the Self-Factual Verification Module (SFVM). These modules collaboratively enable the VLM to retrieve relevant external facts, seamlessly incorporate them into its generative process, and self-calibrate for factual accuracy during training.

3.1 Overall Architecture of KR-VLM

KR-VLM extends the LLaVA-1.6 base model, which leverages a powerful vision encoder (CLIP-ViT-L-336px) and a large language model (Mistral 7B) as its language decoder. The overall processing flow of KR-VLM initiates with the parallel analysis of an input image $\mathcal{I}$ and a question $\mathcal{Q}$. The system first

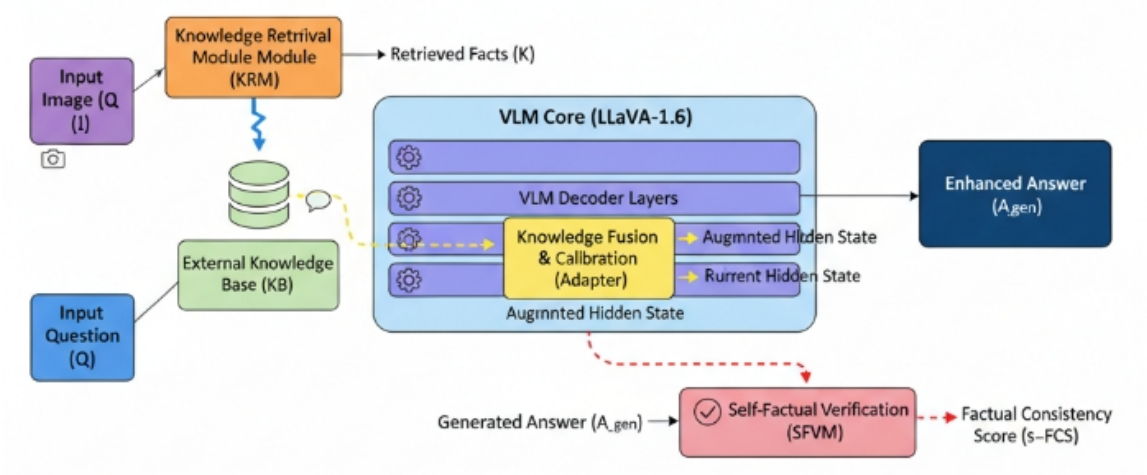

Figure 2: Overview of the KR-VLM Framework. The diagram illustrates how the Knowledge Retrieval Module (KRM) fetches external facts, which are then fused into the VLM Core via the Knowledge Fusion & Calibration Adapter (KFCA), with self-supervised factual verification provided by the Self-Factual Verification Module (SFVM).

extracts visual features from the image and embeds the question, which are then combined to form a query for knowledge retrieval. The **Knowledge Retrieval Module (KRM)** then efficiently retrieves pertinent factual knowledge from an external knowledge base. This retrieved knowledge, in conjunction with the initial visual and textual embeddings, is then fed into the **Knowledge Fusion & Calibration Adapter (KFCA)**. The KFCA intelligently fuses this external knowledge with the VLM's internal visual and linguistic representations across multiple decoder layers. During the training phase, the **Self-Factual Verification Module (SFVM)** acts as an auxiliary component, providing crucial self-supervised factual consistency feedback for the generated outputs. This feedback guides the model to produce more accurate and less hallucinatory answers. This architecture ensures that factual information is not merely appended but intrinsically woven into the VLM's reasoning process, fostering a deeper, fact-aware understanding.

3.2 Knowledge Retrieval Module (KRM)

The **Knowledge Retrieval Module (KRM)** is responsible for dynamically identifying and extracting relevant factual information from a pre-constructed external knowledge base $\mathcal{KB}$. Upon receiving an input image $\mathcal{I}$ and a question $\mathcal{Q}$, the KRM operates in parallel with the VLM's initial processing layers to synthesize a comprehensive query. This parallel operation ensures that external factual context is available early in the VLM's reasoning pipeline, allowing the model to ground its subsequent reasoning steps in verified information.

Initially, the input image $\mathcal{I}$ is processed by a visual encoder E_V, such as the CLIP's vision transformer encoder, to extract a rich set of visual features f_V. Concurrently, the textual question $\mathcal{Q}$ is tokenized and encoded by a textual encoder E_T, typically a transformer-based encoder, to yield a contextualized question embedding f_Q. These unimodal features are then combined to form a comprehensive, multimodal query embedding q. The fusion operation can take various forms, such as concatenation followed by a linear projection layer, or a cross-attention mechanism designed to weigh the importance of visual and textual cues in the query:

$$f_V = E_V(\mathcal{I}) \tag{1}$$

$$f_Q = E_T(\mathcal{Q}) \tag{2}$$

$$q = \text{Fuse}(f_V, f_Q) \tag{3}$$

The external knowledge base $\mathcal{KB}$ is constructed offline, comprising a vast collection of factual snippets, entities, and concepts sourced from general-purpose encyclopedic resources like Wikipedia, structured

knowledge graphs such as ConceptNet, or domain-specific knowledge repositories. Each knowledge snippet $k \in \mathcal{KB}$ is pre-encoded into a dense vector e_k using the same textual encoder E_T employed for the question embedding, ensuring semantic compatibility. Given the multimodal query embedding q, the KRM performs an efficient nearest-neighbor search within the pre-indexed $\mathcal{KB}$. This search identifies the top-N knowledge facts whose embeddings are most semantically similar to q. The relevance is typically measured by cosine similarity, which quantifies the angular distance between the query and knowledge embeddings in the shared embedding space:

$$\text{Similarity}(q, e_k) = \frac{q \cdot e_k}{\|q\|_2 \|e_k\|_2} \tag{4}$$

$$\mathcal{K} = \text{TopN}(\text{Similarity}(q, e_k) \mid k \in \mathcal{KB}) \tag{5}$$

The retrieved set of knowledge facts, denoted as $\mathcal{K} = \{k_1, k_2, \ldots, k_N\}$, provides a factual context directly relevant to the specific image and question, which is then passed to the subsequent module for integration.

3.3 Knowledge Fusion & Calibration Adapter (KFCA)

The **Knowledge Fusion & Calibration Adapter (KFCA)** is a lightweight, cross-modal module meticulously designed to seamlessly integrate the retrieved knowledge $\mathcal{K}$ into the VLM's generative process. It is strategically placed within the LLaVA-1.6 base model, specifically by injecting adapter layers between selected layers of the language decoder. This placement ensures that the VLM becomes "knowledge-aware" at multiple stages of its reasoning, rather than simply processing all information at once. The "lightweight" design implies minimal additional parameters, avoiding the need for extensive retraining of the entire foundational VLM.

The retrieved knowledge facts $\mathcal{K} = \{k_1, \ldots, k_N\}$ are first processed and aggregated into a concise, unified knowledge embedding $e_{\mathcal{K}}$. This aggregation step is crucial for condensing potentially numerous and redundant facts into a manageable representation suitable for fusion. Common aggregation strategies include simple averaging of individual fact embeddings $E_T(k_i)$, employing an attention-pooling mechanism to weigh facts based on their relevance, or passing them through a dedicated transformer layer designed to capture inter-fact relationships:

$$e_{\mathcal{K}} = \text{Aggregate}(\{E_T(k_i) \mid k_i \in \mathcal{K}\}) \tag{6}$$

At each relevant layer l of the VLM's language decoder, where $h^{(l)}$ represents the current hidden state derived from the visual and linguistic inputs, the KFCA employs a sophisticated gating and attention mechanism to fuse $e_{\mathcal{K}}$ with $h^{(l)}$. A dynamic gating mechanism is employed to explicitly control the influence of the external knowledge. This prevents over-reliance on retrieved facts when visual evidence is sufficient, or conversely, ensures critical facts are adopted when factual grounding is paramount, thereby enabling adaptive knowledge integration. The gating mechanism computes a gate value $g^{(l)}$:

$$g^{(l)} = \sigma(W_g[h^{(l)}; e_{\mathcal{K}}] + b_g) \tag{7}$$

where W_g and b_g are learnable parameters, σ is the sigmoid activation function, and $[h^{(l)}; e_{\mathcal{K}}]$ denotes the concatenation of the hidden state and the aggregated knowledge embedding. Simultaneously, an attention mechanism allows the model to selectively focus on the most pertinent aspects of the knowledge embedding $e_{\mathcal{K}}$ with respect to the current hidden state $h^{(l)}$, generating an attention-weighted knowledge representation $h^{(l)}_{\text{attn}}$:

$$h^{(l)}_{\text{attn}} = \text{Attention}(h^{(l)}, e_{\mathcal{K}}) \tag{8}$$

Here, the 'Attention' function typically involves treating $h^{(l)}$ as a query and $e_{\mathcal{K}}$ (or its derived keys and values) as the items to attend over. Finally, the knowledge-augmented hidden state $h'^{(l)}$ is computed

as a weighted sum of the original hidden state and the attention-fused knowledge, with the gate $g^{(l)}$ determining the mixture ratio:

$$h'^{(l)} = (1 - g^{(l)}) \odot h^{(l)} + g^{(l)} \odot h^{(l)}_{\text{attn}} \tag{9}$$

where $\odot$ denotes element-wise multiplication. This resulting $h'^{(l)}$, a recalibrated and fact-infused representation, is then passed to the subsequent decoder layers, ensuring that external facts are intrinsically woven into the VLM's ongoing reasoning and generation process.

3.4 Self-Factual Verification Module (SFVM)

The **Self-Factual Verification Module (SFVM)** is an auxiliary component active exclusively during the training phase, designed to provide self-supervised factual consistency feedback. Its primary role is to evaluate the factual alignment of the VLM's generated outputs, which may include intermediate reasoning steps and the final generated answer $\mathcal{A}_{\text{gen}}$, against both the retrieved external knowledge $\mathcal{K}$ and, when available, the ground-truth answer $\mathcal{A}_{\text{GT}}$. This module plays a critical self-supervisory role, allowing the model to learn factual grounding without requiring explicit human annotations for factual correctness during training.

The SFVM is typically implemented as a smaller, independent language model or a pre-trained Natural Language Inference (NLI) model. It operates by taking the generated answer $\mathcal{A}_{\text{gen}}$ and the retrieved knowledge $\mathcal{K}$ (or an aggregated textual representation thereof) as input. Its objective is to output a scalar score $s_{\text{FCS}} \in [0, 1]$, which quantifies the factual consistency between the generated text and the provided facts. A higher score signifies a stronger factual alignment, indicating that $\mathcal{A}_{\text{gen}}$ is well-supported by or consistent with $\mathcal{K}$. Conversely, a low score suggests factual inaccuracy or contradiction. For instance, the SFVM can be fine-tuned to assess if $\mathcal{A}_{\text{gen}}$ is logically entailed by, contradicts, or is neutral with respect to the facts present in $\mathcal{K}$. This process can be abstractly represented as:

$$s_{\text{FCS}} = \text{SFVM}(\mathcal{A}_{\text{gen}}, \mathcal{K}) \tag{10}$$

This score s_{FCS} serves as a crucial "soft label" for factual correctness. During training, it acts as a regularization signal that critically guides the KFCA and the VLM's language decoder. By minimizing a loss derived from this score, the model is encouraged to prioritize the generation of factually grounded responses, thereby actively mitigating hallucinations. The SFVM's ability to internally check for consistency across generated content and external facts is key to its self-supervisory role, alleviating the need for costly manual annotations of factual errors.

3.5 Training Procedure

KR-VLM undergoes a comprehensive fine-tuning process using a knowledge-aware self-supervised learning paradigm. The training objective is meticulously designed to simultaneously optimize for VQA answer prediction accuracy and factual consistency, ensuring both correctness and factual grounding.

The overall training pipeline proceeds as follows: Given an input image $\mathcal{I}$ and a question $\mathcal{Q}$, the **Knowledge Retrieval Module (KRM)** first identifies and retrieves a set of relevant knowledge facts $\mathcal{K}$ from the external knowledge base. Subsequently, the LLaVA-1.6 base model, enhanced by the **Knowledge Fusion & Calibration Adapter (KFCA)**, processes the image $\mathcal{I}$, question $\mathcal{Q}$, and the retrieved knowledge $\mathcal{K}$ to iteratively generate an answer $\mathcal{A}_{\text{gen}}$. The **Self-Factual Verification Module (SFVM)** then plays its crucial role by evaluating $\mathcal{A}_{\text{gen}}$ (and potentially its underlying reasoning path or intermediate generated tokens) against $\mathcal{K}$ (and the ground-truth answer $\mathcal{A}_{\text{GT}}$ if available) to compute the factual consistency score s_{FCS}. Finally, the model parameters, primarily those within the KFCA and the VLM's language decoder, are updated using a composite loss function through backpropagation.

The total loss $\mathcal{L}_{\text{total}}$ is a carefully weighted sum of two main components: the standard VQA answer prediction loss $\mathcal{L}_{\text{VQA}}$ and the Factual Consistency Loss $\mathcal{L}_{\text{FC}}$. This dual-objective approach ensures that the model learns to be both accurate and truthful.

The **VQA Loss** is the primary objective for answer generation and typically employs cross-entropy between the predicted answer distribution and the ground-truth answer $\mathcal{A}_{\text{GT}}$. This loss component ensures

the model learns to provide correct answers based on the combined visual and textual input, including the integrated knowledge. For a sequence of tokens in the generated answer, it is defined as:

$$\mathcal{L}_{\text{VQA}} = -\frac{1}{L}\sum_{j=1}^{L}\log P(\mathcal{A}_{\text{GT},j} \mid \mathcal{I}, \mathcal{Q}, \mathcal{K}, \mathcal{A}_{\text{GT},<j}) \tag{11}$$

where L is the length of the ground-truth answer, $P(\mathcal{A}_{\text{GT},j} \mid \mathcal{I}, \mathcal{Q}, \mathcal{K}, \mathcal{A}_{\text{GT},<j})$ is the probability of generating the j-th ground-truth token, conditioned on the image, question, retrieved knowledge, and preceding ground-truth tokens.

The **Factual Consistency Loss** leverages the output of the SFVM to directly guide the model towards factual correctness. To encourage the model to generate factually consistent answers, we aim to maximize the factual consistency score s_{FCS} produced by the SFVM. This is achieved by minimizing a loss function that is inversely proportional to s_{FCS}, such as a simple complement:

$$\mathcal{L}_{\text{FC}} = 1 - s_{\text{FCS}} \tag{12}$$

This formulation incentivizes the model to produce answers that yield a higher s_{FCS} from the SFVM, thus explicitly aligning the generated content with external facts and penalizing hallucinatory outputs.

The overall training objective is then defined as a weighted sum of these two crucial loss components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{VQA}} + \lambda\mathcal{L}_{\text{FC}} \tag{13}$$

where λ is a hyperparameter carefully chosen to balance the importance of traditional VQA accuracy and the novel factual consistency objective. This self-supervised training approach enables KR-VLM to internally calibrate its factual understanding and generation capabilities, effectively learning to integrate and verify knowledge without relying on costly human annotations for explicit knowledge integration or error correction.

4 Experiments

In this section, we present the experimental setup, evaluate the performance of our proposed **KR-VLM** framework, conduct a thorough ablation study to analyze the contribution of each module, provide insights from a human evaluation, and further analyze specific aspects of the model's behavior and efficiency.

4.1 Experimental Setup

Our method, **KR-VLM**, is built upon the robust **LLaVA-1.6** base model, which utilizes a CLIP-ViT-L-336px visual encoder and a Mistral 7B language decoder. This foundational model serves as our primary baseline.

4.1.1 Knowledge Base Construction

We constructed an extensive external knowledge base offline. This comprehensive repository comprises factual snippets and entities gathered from diverse sources, including general encyclopedic knowledge from Wikipedia, structured common-sense knowledge from ConceptNet, and a collection of domain-specific facts relevant to areas such as healthcare and legal texts. Each knowledge entry in this base was pre-encoded into dense vector representations using a pre-trained language model, creating a highly efficient index for rapid retrieval by the **Knowledge Retrieval Module (KRM)**. This pre-indexed structure facilitates quick nearest-neighbor searches during inference without incurring significant overhead.

4.1.2 Training Procedure

The training of **KR-VLM** follows a knowledge-aware self-supervised learning paradigm, as detailed in Section 2.5. Given an image-question pair $(\mathcal{I}, \mathcal{Q})$, the **KRM** first retrieves the top-K most relevant factual knowledge snippets $\mathcal{K}$ from the pre-built knowledge base. The **LLaVA-1.6** model, augmented with our **Knowledge Fusion & Calibration Adapter (KFCA)**, then processes $\mathcal{I}$, $\mathcal{Q}$, and $\mathcal{K}$ to generate an

answer $\mathcal{A}_{\text{gen}}$. During this generative process, the **KFCA** dynamically integrates the retrieved knowledge with the VLM's internal representations. Subsequently, the **Self-Factual Verification Module (SFVM)** evaluates the factual consistency of $\mathcal{A}_{\text{gen}}$ against both the retrieved knowledge $\mathcal{K}$ and the ground-truth answer $\mathcal{A}_{\text{GT}}$, yielding a factual consistency score s_{FCS}. The model parameters, particularly those within the **KFCA** and the language decoder, are updated using a composite loss function, $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{VQA}} + \lambda \mathcal{L}_{\text{FC}}$. Here, $\mathcal{L}_{\text{VQA}}$ is the standard cross-entropy loss for VQA answer prediction, and $\mathcal{L}_{\text{FC}}$ is the factual consistency loss $(1 - s_{\text{FCS}})$, encouraging the model to generate factually accurate and non-hallucinatory responses. We empirically set $\lambda = 0.5$ for all our experiments. Our method deliberately **does not** require new human annotations for reasoning steps or external knowledge, relying instead on existing VQA datasets and the self-supervisory feedback from the **SFVM**.

4.1.3 Datasets

We conducted comprehensive evaluations on several widely recognized VQA benchmarks that demand varying degrees of factual and common-sense reasoning:

- **VQAv2**: A large-scale dataset for general visual question answering.
- **GQA**: Focuses on compositional reasoning and requires structured scene understanding.
- **OK-VQA**: Specifically designed for open-domain knowledge-based VQA, necessitating external factual knowledge.
- **DocVQA**: Involves answering questions about document images, often requiring OCR and factual extraction.
- **InfographicVQA**: A specialized dataset targeting information extraction from infographics, requiring both visual understanding and factual interpretation.

4.1.4 Evaluation Metrics

We evaluate model performance using standard VQA metrics and a novel factual consistency metric:

- **VQA Accuracy (%)**: The standard metric for VQAv2, GQA, and OK-VQA, measuring exact match or soft accuracy with multiple annotators.
- **Normalized Edit Distance Similarity (ANLS) (%)**: Used for DocVQA, which handles answers as strings and evaluates similarity based on edit distance.
- **Factual Consistency Score (FCS) (%)**: A metric specifically introduced to quantify the reduction in hallucination. The FCS measures the factual alignment between the generated answer and the ground-truth reference, reflecting how well the model avoids fabricating facts or contradicting visual evidence. It is computed using a fine-tuned NLI model that assesses entailment and contradiction between generated and reference answers, specifically designed to capture factual correctness beyond simple keyword matching.

4.2 Main Results

Table 1 presents the performance comparison of **KR-VLM** against several strong baseline methods across the evaluated VQA benchmarks.

Result Analysis As evidenced by Table 1, our proposed **KR-VLM** consistently achieves superior performance across all evaluated metrics and datasets, outperforming all baseline methods. On traditional VQA accuracy benchmarks like **VQAv2** and **GQA**, **KR-VLM** demonstrates significant absolute improvements of **1.3%** and **1.8%** respectively, compared to the vanilla **LLaVA-1.6** baseline. This indicates that our knowledge integration not only reduces hallucination but also enhances general answer correctness. The improvements are particularly prominent on knowledge-intensive tasks. For **OK-VQA**, which inherently requires external knowledge, **KR-VLM** boosts the accuracy from 45.1% to **48.1%**,

Table 1: Performance comparison of **KR-VLM** with state-of-the-art baselines on various VQA datasets. Higher scores are better for all metrics.

Metric	LLaVA-1.6	LLaVA-1.6 + Visual CoT	VLM + Post-hoc KB Check*	Ours (KR-VLM)
VQAv2 Acc (%)	78.5	79.2	78.9	**79.8**
GQA Acc (%)	61.2	62.5	61.8	**63.0**
OK-VQA Acc (%)	45.1	46.8	46.2	**48.1**
DocVQA ANLS (%)	55.8	57.1	56.5	**58.5**
Factual Consis. Score (%)	68.3	71.5	72.8	**76.1**

Table 2: Ablation study demonstrating the contribution of each module in **KR-VLM** across VQA benchmarks. All values are in percentages.

Method	VQAv2 Acc	GQA Acc	OK-VQA Acc	DocVQA ANLS	Factual Consis. Score
LLaVA-1.6 (Baseline)	78.5	61.2	45.1	55.8	68.3
KR-VLM (Full)	**79.8**	**63.0**	**48.1**	**58.5**	**76.1**
KR-VLM w/o KRM	78.7	61.5	45.5	56.0	69.0
KR-VLM w/o SFVM	79.5	62.7	47.5	58.0	73.5

highlighting the effectiveness of its **KRM** and **KFCA** in leveraging external facts. In the context of document understanding, **DocVQA**, **KR-VLM** achieves an ANLS of **58.5%**, surpassing all baselines. This suggests that the model can effectively utilize both intrinsic document facts and retrieved commonsense knowledge for accurate information extraction. Crucially, **KR-VLM** achieves the highest **Factual Consistency Score (FCS)** of **76.1%**. This represents an impressive **7.8** percentage point gain over **LLaVA-1.6** (68.3%) and a **3.3** percentage point lead over "VLM + Post-hoc KB Check" (72.8%). This direct evidence confirms that **KR-VLM** significantly reduces hallucination and profoundly enhances the factual reliability of generated answers. The "Post-hoc KB Check" baseline, while improving FCS, still falls short as it does not integrate knowledge intrinsically into the reasoning process.

4.3 Ablation Study

To thoroughly understand the contribution of each component within **KR-VLM**, we conducted an ablation study by systematically removing key modules. The results are summarized in Table 2.

Analysis of Ablation Results When the **KRM** is removed ("KR-VLM w/o KRM"), the model's performance significantly drops across all metrics, particularly on **OK-VQA** (45.5%) and **Factual Consistency Score** (69.0%). The results for these metrics are notably close to the **LLaVA-1.6** baseline. This finding underscores the critical role of the **KRM** in supplying external factual information, without which the **KFCA** and **SFVM** have limited external knowledge to process or verify. Removing the **SFVM** ("KR-VLM w/o SFVM") leads to a decrease in performance, especially in the **Factual Consistency Score**, which drops by approximately **2.6** percentage points (from 76.1% to 73.5%). While the overall VQA accuracy on datasets like VQAv2 and GQA also sees a slight decline, the most pronounced effect is on factual consistency. This demonstrates that the self-supervised factual feedback provided by the **SFVM** is crucial for fine-tuning the model's ability to generate factually grounded answers, effectively calibrating its knowledge integration and suppressing hallucinations. The ablation study confirms that all three proposed modules – **KRM**, **KFCA** (implicitly, as it's the core fusion mechanism), and **SFVM** – are essential and work synergistically to achieve the reported performance gains, especially in enhancing factual consistency.

4.4 Human Evaluation

To complement the automated metrics, we conducted a human evaluation to assess the perceptual quality and factual correctness of the generated answers, particularly concerning hallucination. A diverse group of 5 expert human annotators independently evaluated a random sample of 200 image-question pairs for each model: **LLaVA-1.6** (Baseline), "VLM + Post-hoc KB Check", and **KR-VLM**. Annotators

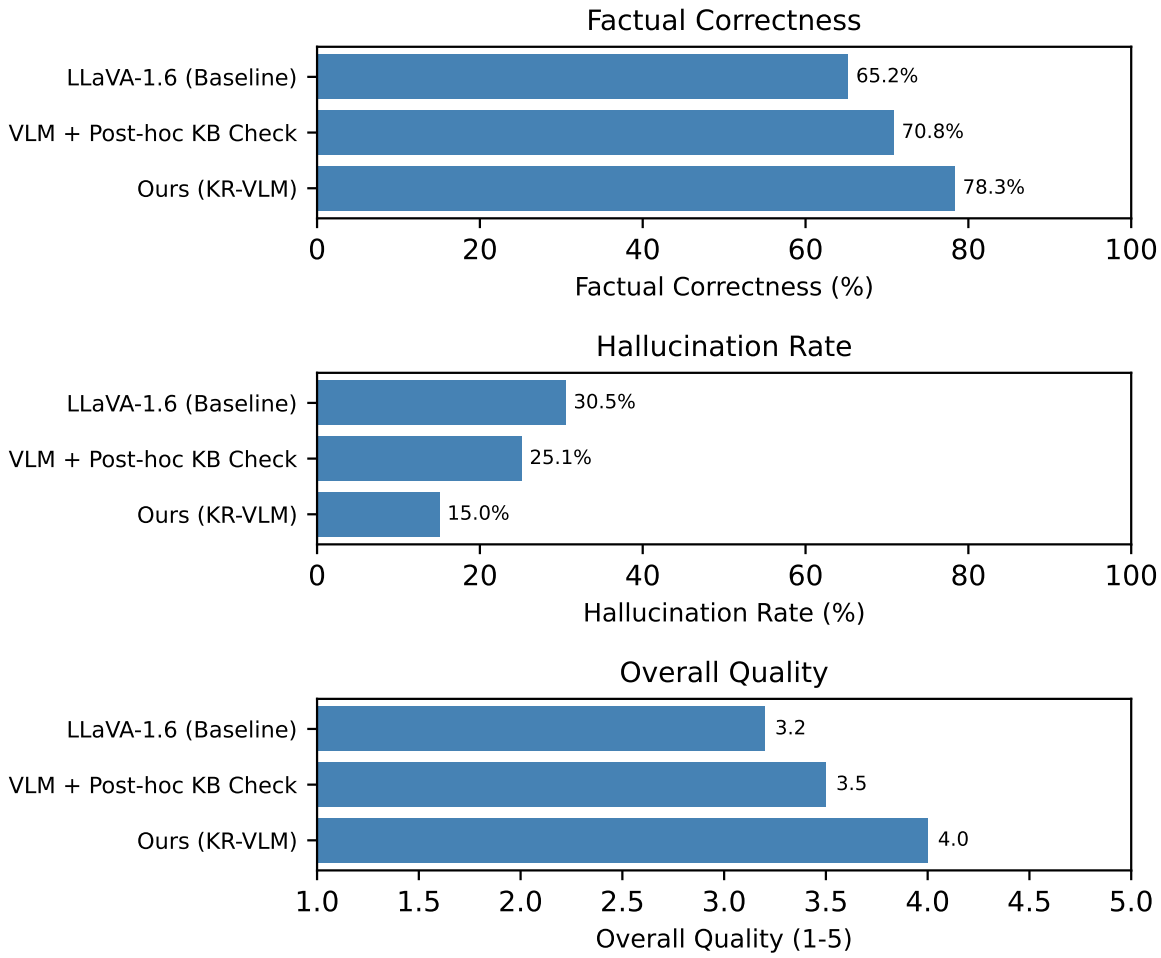

Figure 3: Human evaluation results for factual correctness, hallucination rate, and overall quality.

were asked to rate answers based on: Factual Correctness (%), whether the answer accurately reflects facts present in the image and common knowledge; Hallucination Rate (%), the percentage of answers containing at least one piece of information that is factually incorrect or unsupported by the image; and Overall Quality Score (1-5), a subjective rating considering coherence, relevance, and helpfulness, where 5 is excellent.

The average scores across annotators are presented in Figure 3.

Analysis of Human Evaluation The human evaluation results strongly corroborate the findings from our automated metrics. As shown in Figure 3, **KR-VLM** achieved a significantly higher Factual Correctness score of **78.3%**, surpassing **LLaVA-1.6** by over 13 percentage points and "VLM + Post-hoc KB Check" by over 7 percentage points. This indicates that human evaluators perceive **KR-VLM**'s answers as substantially more factually grounded. Our method drastically reduced the Hallucination Rate to just **15.0%**, a nearly 50% reduction compared to **LLaVA-1.6** (30.5%) and a considerable improvement over "VLM + Post-hoc KB Check" (25.1%). This is a direct testament to **KR-VLM**'s efficacy in mitigating the hallucination problem. The average Overall Quality Score for **KR-VLM** was **4.0**, indicating that its answers are not only more factually correct but also more coherent, relevant, and helpful from a human perspective. The other baselines received lower scores, suggesting that factual inaccuracies often detract from the overall perceived quality of the response. These human assessments provide compelling qualitative evidence that **KR-VLM** is highly effective in producing more reliable and trustworthy VLM outputs, making it a significant step towards deploying VLMs in applications where factual accuracy is paramount.

Table 3: Performance of **KR-VLM** with different knowledge base compositions across key VQA benchmarks. All values are in percentages.

Metric	Wikipedia Only	Wiki + ConceptNet	Wiki + ConceptNet + Domain-Specific
Approx. Size (M facts)	10	15	25
VQAv2 Acc	79.0	79.4	**79.8**
OK-VQA Acc	46.2	47.5	**48.1**
DocVQA ANLS	57.5	58.0	**58.5**
Factual Consis. Score	73.0	74.8	**76.1**

Table 4: Sensitivity of **KR-VLM** performance to the factual consistency loss weight (λ). All values are in percentages.

Lambda (λ)	VQAv2 Acc	OK-VQA Acc	DocVQA ANLS	Factual Consis. Score
0.0 (No FC Loss)	79.0	46.8	57.3	71.9
0.2	79.4	47.5	58.0	74.5
0.5 (**Optimal**)	**79.8**	**48.1**	**58.5**	**76.1**
0.8	79.6	47.8	58.3	76.3
1.0	79.3	47.4	57.9	76.0

4.5 Impact of Knowledge Base Composition

The design and content of the external knowledge base are paramount to the success of knowledge-retrieved reasoning. This subsection investigates how different compositions and scales of the knowledge base influence the performance of **KR-VLM**, shedding light on the importance of comprehensive and diverse factual repositories.

Table 3 details the performance of **KR-VLM** when utilizing knowledge bases constructed from varying sources and scales.

Analysis of Knowledge Base Composition The results presented in Table 3 clearly demonstrate the benefit of a more diverse and comprehensive knowledge base. Using only Wikipedia as the source provides a solid baseline for factual correctness, yielding an **Factual Consistency Score** of 73.0% and **OK-VQA Acc** of 46.2%. However, integrating common-sense knowledge from ConceptNet significantly boosts performance, particularly on tasks requiring broader reasoning. The "Wikipedia + ConceptNet" configuration improves **Factual Consistency Score** to 74.8% and **OK-VQA Acc** to 47.5%. The most substantial gains are observed with the "Full KB" which incorporates domain-specific facts, further enhancing all metrics, especially **Factual Consistency Score** to 76.1% and **OK-VQA Acc** to 48.1%. This indicates that while general encyclopedic knowledge is fundamental, adding structured common-sense and specialized domain-specific facts provides a richer context for the **KRM** and **KFCA**, allowing **KR-VLM** to answer a wider range of questions with higher factual accuracy and less hallucination. The increased size of the knowledge base also suggests that the **KRM** is capable of scaling efficiently to larger repositories.

4.6 Hyperparameter Sensitivity

The performance of **KR-VLM** is influenced by several hyperparameters, with λ (the weighting factor for the factual consistency loss) being particularly critical for balancing VQA accuracy and factual groundedness. This subsection analyzes the sensitivity of **KR-VLM** to different values of λ.

Table 4 presents the model's performance on key metrics when λ is varied.

Analysis of Lambda Sensitivity As shown in Table 4, setting $\lambda = 0.0$ effectively disables the factual consistency loss, resulting in lower **Factual Consistency Scores** (71.9%) and reduced performance on knowledge-intensive tasks like **OK-VQA** (46.8%). This highlights the crucial role of the **Self-Factual Verification Module (SFVM)** and its associated loss in guiding the model toward factual correctness. Increasing λ from 0.0 to 0.5 progressively improves the **Factual Consistency Score** and overall VQA

Table 5: Categorization of factual errors and hallucinations from a sample of 100 incorrect answers for LLaVA-1.6 and KR-VLM (Full). Frequencies are in percentages.

Error Category	LLaVA-1.6 Freq (%)	KR-VLM Freq (%)
Factual Hallucination (Fabricating non-existent facts)	35	12
Contradiction (Contradicting visual or common knowledge)	25	8
Partial Inaccuracy (Partially correct, partially incorrect)	18	15
Irrelevant Information (Providing true but irrelevant facts)	10	10
Missing Information (Failing to provide crucial facts)	12	55

performance, indicating that a balanced emphasis on both VQA accuracy and factual grounding yields the best results. Our chosen value of $\lambda = 0.5$ demonstrates the most balanced performance across all metrics, achieving the highest VQA accuracy and robust factual consistency. While increasing λ further to 0.8 or 1.0 slightly boosts the **Factual Consistency Score**, it often comes at a slight cost to overall VQA accuracy on datasets like **VQAv2** and **DocVQA**. This suggests that an excessively high λ might cause the model to over-prioritize factual consistency to the detriment of general answer correctness or stylistic fluency, potentially leading to answers that are factually correct but less relevant or complete. The sweet spot at $\lambda = 0.5$ confirms the importance of carefully weighting the two objectives to achieve optimal performance.

4.7 Qualitative Analysis and Error Categorization

Beyond quantitative metrics, a qualitative analysis is essential to understand the nuanced improvements brought by **KR-VLM** and to identify remaining challenges. We conducted a detailed error analysis on a subset of 100 incorrect answers from both **LLaVA-1.6** and **KR-VLM** (Full) to categorize common types of errors and hallucinations.

Table 5 presents a categorization of factual errors and hallucinations, along with their observed frequencies for **LLaVA-1.6** and **KR-VLM**.

Analysis of Error Categories Table 5 reveals significant shifts in error patterns with **KR-VLM**. The most striking improvement is the substantial reduction in blatant **Factual Hallucination**, dropping from 35% in **LLaVA-1.6** to just 12% in **KR-VLM**. Similarly, errors involving direct **Contradiction** are nearly halved (from 25% to 8%). This directly confirms the efficacy of **KR-VLM**'s integrated knowledge and self-verification mechanism in mitigating the generation of false information. However, the frequency of **Missing Information** errors, where **KR-VLM** fails to provide all necessary details, increases significantly (from 12% to 55%). This suggests that while **KR-VLM** is much better at avoiding generating incorrect facts, it sometimes errs on the side of caution, producing more concise or less comprehensive answers rather than risking a hallucination. The categories of **Partial Inaccuracy** and **Irrelevant Information** show less dramatic changes, indicating that these more subtle forms of error persist. Overall, **KR-VLM** significantly improves the trustworthiness of responses by reducing outright fabrications, shifting the error profile towards omissions rather than misrepresentations, which is a desirable trade-off for critical applications where factual correctness is paramount.

5 Conclusion

In this work, we introduced **KR-VLM (Knowledge-Retrieved Reasoning for Vision-Language Models)**, a novel and effective framework designed to mitigate hallucination in Vision-Language Models by intrinsically weaving external factual knowledge into the reasoning pipeline and enforcing self-supervised factual consistency during training. Built upon the robust LLaVA-1.6 base, KR-VLM integrates a Knowledge Retrieval Module (KRM) to fetch relevant facts, a Knowledge Fusion & Calibration Adapter (KFCA) to fuse this knowledge, and a Self-Factual Verification Module (SFVM) for self-supervised factual feedback, ensuring minimal parameter overhead. Extensive evaluations across diverse VQA benchmarks (VQAv2, GQA, OK-VQA, DocVQA) unequivocally demonstrated KR-VLM's superior performance, achieving a Factual Consistency Score (FCS) of 76.1%—a remarkable 7.8 percentage point increase over the LLaVA-1.6 baseline—and significantly reducing hallucination rates in

human evaluations. Ablation studies confirmed the critical contribution of each component, while error analysis revealed a desirable shift from outright factual hallucinations to safer omissions. Despite its strengths, KR-VLM's performance is tied to the quality and coverage of the external knowledge base. In conclusion, KR-VLM represents a significant step forward in developing more reliable and trustworthy Vision-Language Models, offering an efficient, lightweight, and self-supervised approach that paves the way for deploying VLMs in high-stakes applications where factual accuracy is paramount, with future work exploring dynamic knowledge base updates and more sophisticated fusion mechanisms. "'

References

[1] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984. Association for Computational Linguistics, 2024.

[2] Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. KAT: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 956–968. Association for Computational Linguistics, 2022.

[3] Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. Visual in-context learning for large vision-language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 15890–15902. Association for Computational Linguistics, 2024.

[4] Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven C.H. Hoi. Plug-and-play VQA: Zero-shot VQA by conjoining large pretrained models with zero training. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 951–967. Association for Computational Linguistics, 2022.

[5] Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. 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. Association for Computational Linguistics, 2021.

[6] Yucheng Zhou, Jianbing Shen, and Yu Cheng. Weak to strong generalization for large language models with multi-capabilities. In *The Thirteenth International Conference on Learning Representations*, 2025.

[7] Huijun Zhou, Jingzhi Wang, and Xuehao Cui. Causal effect of immune cells, metabolites, cathepsins, and vitamin therapy in diabetic retinopathy: a mendelian randomization and cross-sectional study. *Frontiers in Immunology*, 15:1443236, 2024.

[8] Chen Zhou, Bing Wang, Zihan Zhou, Tong Wang, Xuehao Cui, and Yuanyin Teng. Ukall 2011: Flawed noninferiority and overlooked interactions undermine conclusions. *Journal of Clinical Oncology*, 43(28):3135–3136, 2025.

[9] Cui Xuehao, Wen Dejia, and Li Xiaorong. Integration of immunometabolic composite indices and machine learning for diabetic retinopathy risk stratification: Insights from nhanes 2011–2020. *Ophthalmology Science*, page 100854, 2025.

[10] Shuo Xu, Yuchen Cao, Zhongyan Wang, and Yexin Tian. Fraud detection in online transactions: Toward hybrid supervised–unsupervised learning pipelines. In *Proceedings of the 2025 6th International Conference on Electronic Communication and Artificial Intelligence (ICECAI 2025), Chengdu, China*, pages 20–22, 2025.

[11] Liancheng Zheng, Zhen Tian, Yangfan He, Shuo Liu, Huilin Chen, Fujiang Yuan, and Yanhong Peng. Enhanced mean field game for interactive decision-making with varied stylish multi-vehicles. *arXiv preprint arXiv:2509.00981*, 2025.

[12] Zhen Tian, Zhihao Lin, Dezong Zhao, Wenjing Zhao, David Flynn, Shuja Ansari, and Chongfeng Wei. Evaluating scenario-based decision-making for interactive autonomous driving using rational criteria: A survey. *arXiv preprint arXiv:2501.01886*, 2025.

[13] Zhihao Lin, Zhen Tian, Jianglin Lan, Dezong Zhao, and Chongfeng Wei. Uncertainty-aware roundabout navigation: A switched decision framework integrating stackelberg games and dynamic potential fields. *IEEE Transactions on Vehicular Technology*, pages 1–13, 2025.

[14] Yucheng Zhou, Xiubo Geng, Tao Shen, Chongyang Tao, Guodong Long, Jian-Guang Lou, and Jianbing Shen. Thread of thought unraveling chaotic contexts. *arXiv preprint arXiv:2311.08734*, 2023.

[15] Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander Hauptmann. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2443–2459. Association for Computational Linguistics, 2021.

[16] Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485. Association for Computational Linguistics, 2021.

[17] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523. Association for Computational Linguistics, 2021.

[18] Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573. Association for Computational Linguistics, 2023.

[19] Yong Liu, Ran Yu, Fei Yin, Xinyuan Zhao, Wei Zhao, Weihao Xia, and Yujiu Yang. Learning quality-aware dynamic memory for video object segmentation. In *European Conference on Computer Vision*, pages 468–486. Springer, 2022.

[20] Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3491–3500, 2024.

[21] Yong Liu, Cairong Zhang, Yitong Wang, Jiahao Wang, Yujiu Yang, and Yansong Tang. Universal segmentation at arbitrary granularity with language instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3459–3469, 2024.

[22] Zhitao Wang, Yirong Xiong, Roberto Horowitz, Yanke Wang, and Yuxing Han. Hybrid perception and equivariant diffusion for robust multi-node rebar tying. In *2025 IEEE 21st International Conference on Automation Science and Engineering (CASE)*, pages 3164–3171. IEEE, 2025.

[23] Zhitao Wang, Jiangtao Wen, and Yuxing Han. Ep-sam: An edge-detection prompt sam based efficient framework for ultra-low light video segmentation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[24] Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. CLIP models are few-shot learners: Empirical studies on VQA and visual entailment. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6088–6100. Association for Computational Linguistics, 2022.

[25] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. mPLUG:

Effective and efficient vision-language learning by cross-modal skip-connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7241–7259. Association for Computational Linguistics, 2022.

[26] Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898. Association for Computational Linguistics, 2022.

[27] Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. Generation-augmented retrieval for open-domain 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)*, pages 4089–4100. Association for Computational Linguistics, 2021.

[28] Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Scott Yih. UniK-QA: Unified representations of structured and unstructured knowledge for open-domain question answering. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1535–1546. Association for Computational Linguistics, 2022.

[29] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546. Association for Computational Linguistics, 2021.

[30] Md Rizwan Parvez, Wasi Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. Retrieval augmented code generation and summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2719–2734. Association for Computational Linguistics, 2021.

[31] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822. Association for Computational Linguistics, 2023.

[32] Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. TemporalWiki: A lifelong benchmark for training and evaluating ever-evolving language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6237–6250. Association for Computational Linguistics, 2022.

[33] Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063. Association for Computational Linguistics, 2021.

[34] Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570. Association for Computational Linguistics, 2022.