electronics

Article

Bridging Cross-Modal Semantic Gaps with Multi-Source Semantic Anchors in Knowledge-Based Visual Question Answering

Junming Hu [1], Jinxiong Zhang [1,2,3,4], Feng Zhan [1,2,3,4,*] and Yiran Huang [1,2,3,4,*]

[1] School of Computer, Electronics and Information, Guangxi University, Nanning 530004, China; 2313301015@st.gxu.edu.cn (J.H.); zhangjx@gxu.edu.cn (J.Z.)
[2] Guangxi Key Laboratory of Multimedia Communications Network Technology, Nanning 530004, China
[3] Guangxi Intelligent Digital Service Technology Innovation Center, Nanning 530004, China
[4] Key Laboratory of Parallel, Distribution and Intelligent Computing in Guangxi Universities and Colleges, Nanning 530004, China
* Correspondence: zf@gxu.edu.cn (F.Z.); hyr@gxu.edu.cn (Y.H.)

Abstract

Knowledge-based visual question answering (KB-VQA) requires leveraging external knowledge relevant to the image to assist reasoning. Existing methods typically convert images into a single textual description for knowledge retrieval or directly rely on the implicit knowledge within large language models to generate answers. However, a single textual description struggles to preserve fine-grained visual information such as object attributes and scene text, limiting retrieval quality. Meanwhile, naively fusing multi-source information tends to introduce modality noise, undermining reasoning accuracy. To address these issues, we propose a unified framework that constructs multi-source semantic anchors to bridge the cross-modal semantic gaps among vision, questions, and external knowledge. Specifically, we unify image captions, object tags, and optical character recognition (OCR) text as semantic anchors. These anchors serve as shared intermediaries to pre-align visual and textual features, avoiding direct interaction between heterogeneous modalities. During cross-modal fusion, a cross-residual gating mechanism adaptively suppresses modality noise by leveraging the semantic anchors as stable references. The framework further integrates contrastive learning to strengthen cross-modal alignment and employs a retrieve-then-read pipeline for open-domain answer reasoning. Experiments on the OK-VQA, FVQA, and A-OKVQA datasets demonstrate that the proposed framework outperforms state-of-the-art methods across multiple metrics, validating the effectiveness and robustness of the proposed framework.

Keywords: knowledge-based visual question answering; cross-modal alignment; semantic anchors; contrastive learning; dense knowledge retrieval

Check for updates

Academic Editor: Ping-Feng Pai

Received: 30 March 2026
Revised: 20 April 2026
Accepted: 21 April 2026
Published: 26 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Visual Question Answering (VQA), situated at the intersection of natural language processing and computer vision, aims to understand image content and answer related natural language questions [1]. As task complexity increases, traditional VQA methods that rely solely on information within the image can no longer adequately address queries involving common sense or domain-specific knowledge, giving rise to Knowledge-Based VQA (KB-VQA). KB-VQA requires models to incorporate external knowledge bases and

employ explicit or implicit reasoning mechanisms to bridge the cognitive gap between vision and language [2], thereby posing greater challenges for cross-modal understanding and knowledge reasoning. Fundamentally, a natural semantic gap exists between pixel-level visual features and symbol-level linguistic logic, and this gap is further exacerbated by the introduction of external knowledge, as the model must not only align visual and question modalities but also coordinate heterogeneous information from knowledge bases. Therefore, effectively bridging the cross-modal semantic gap and achieving fine-grained fusion of multi-source information remains a central challenge in KB-VQA research.

Existing KB-VQA methods generally follow a two-stage retriever-reader framework, in which a retriever first obtains external knowledge relevant to the image and question, and a reader then predicts the answer. Although prior work [3,4] has made notable progress in leveraging implicit knowledge from large language models or vision–language models, two key limitations persist in multi-modal feature alignment and multi-source information fusion. First, at the feature representation level, existing methods [5,6] typically convert images into a single textual form (e.g., image captions) for retrieval. This practice overlooks crucial fine-grained information in images, such as specific object attributes or scene text, leading to substantial loss of visual semantics during the mapping process and consequently hindering accurate retrieval of relevant knowledge (see Figure 1). This limitation is particularly critical in KB-VQA because, unlike standard VQA, where answers can often be directly inferred from visual content, KB-VQA relies on retrieved knowledge as the primary evidence for reasoning—an imprecise retrieval query therefore propagates errors throughout the entire reasoning pipeline, causing the reader to reason over irrelevant or misleading knowledge and ultimately producing incorrect answers. Second, at the feature fusion level, existing multi-modal interaction mechanisms are generally simplistic: most models rely on straightforward feature concatenation or attention mechanisms and lack effective means to suppress inter-modal noise. In KB-VQA, this deficiency is exacerbated by the introduction of external knowledge as a third information source, which forces the model to simultaneously coordinate three heterogeneous modalities—vision, question, and knowledge—each potentially containing information irrelevant to the current query. Without an explicit mechanism to identify and down-weight such distracting signals during fusion, irrelevant visual regions, ambiguous question tokens, and noisy knowledge entries interfere with each other, degrading the discriminability of the fused representation and ultimately leading to unreliable answer predictions.

To address the aforementioned semantic gap, we present a core insight: image captions, object tags, and Optical Character Recognition (OCR) text all belong to the textual modality and can be directly compared with the question in a unified semantic space; meanwhile, they originate from the image itself and thus inherently encode visual semantic information. Consequently, unifying these three sources as semantic anchors enables them to serve as a bridge that connects the three dimensions of vision, questions, and external knowledge. This idea draws inspiration from prior work in traditional VQA that employs image captions as a cross-modal bridge to close the semantic gap between vision and questions but extends it in two significant ways within the KB-VQA setting. In terms of information coverage, the approach expands from a single caption to a multi-source combination of captions, tags, and OCR text, characterizing image content from three complementary perspectives: global scene semantics, fine-grained entity attributes, and symbolic information. In terms of functionality, semantic anchors serve not only for cross-modal pre-alignment but also simultaneously function as anchoring targets for contrastive learning, reference baselines for residual-gated fusion, and semantic enhancers for knowledge retrieval. Moreover, semantic anchors exhibit a stronger semantic association with visual content than questions do, because anchors directly describe image content,

whereas questions only partially and interrogatively refer to it. Based on this observation, we replace the conventional question–vision pairs with visual–anchor pairs for contrastive learning, thereby more effectively constraining the cross-modal alignment structure of the representation space.

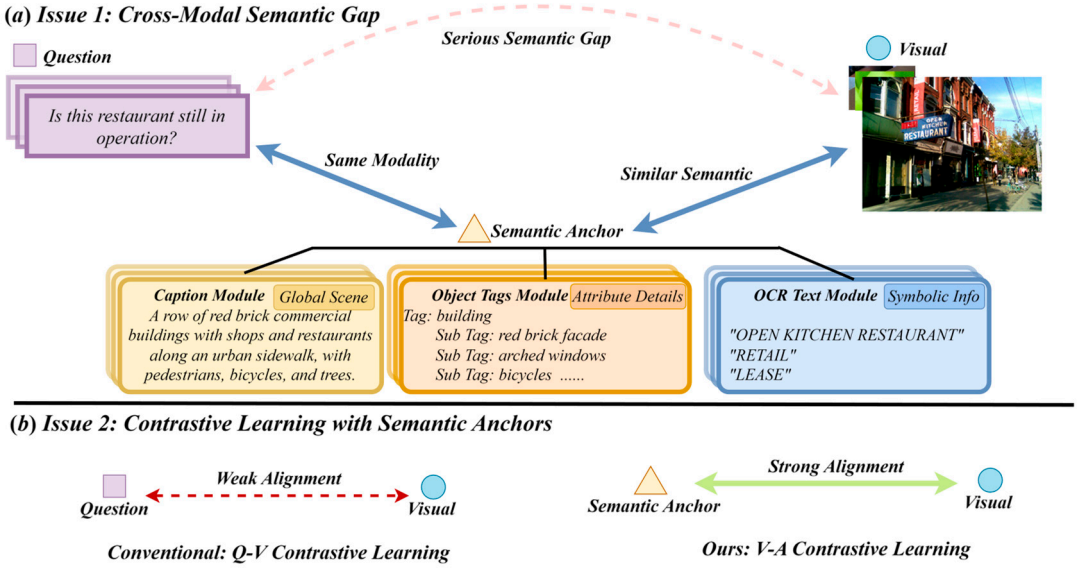


Figure 1. (**a**) A significant semantic gap hinders direct question–visual matching. We construct multi-source semantic anchors—comprising image captions (global scene), object tags (fine-grained attributes), and OCR text (symbolic information)—that share the textual modality with the question while preserving rich semantic associations with the visual content. (**b**) Conventional Q-V contrastive pairs suffer from weak alignment due to the modality gap. Since semantic anchors directly describe image content, V-Sem pairs exhibit stronger semantic connections, enabling more effective alignment of the cross-modal representation space.

Building upon these insights, we propose a unified framework, termed BRIDGE, that constructs multi-source semantic anchors to bridge the cross-modal semantic gaps among vision, questions, and external knowledge. In the feature extraction stage, the Multi-Source Semantic Anchor Construction (MSAC) module integrates image captions, object tags, and OCR text to construct semantic anchors, while a RoBERTa encoder with shared parameters but independent forward passes obtains the representation of each modality. In the cross-modal alignment stage, the Visual–Anchor Cross-Modal Encoder (VACE) and the Question–Anchor Cross-Modal Encoder (QACE) employ semantic anchors as intermediaries to perform symmetric pre-alignment with visual and question features, respectively, thereby establishing a preliminary semantic alignment foundation before deep fusion. In the question–vision (Q-V) fusion stage, the Cross-Residual Gated Q-V Fusion (CRGF) module embeds a cross-residual gated modulation mechanism on top of a standard cross-modal encoder block. It treats semantic anchors as a fixed reference and leverages semantic residuals from complementary modalities to drive gating signals that adaptively suppress modality noise, while Variational Information Bottleneck (VIB) regularization constrains the information efficiency of the fused representation [7]. During training, the Visual–Anchor Contrastive Learning (VACL) module introduces a cross-modal contrastive learning objective that directly constrains the alignment quality between visual features and semantic anchors in a shared space. In the inference and generation stage, the Dense Knowledge Retriever (DKR) retrieves relevant external knowledge based on the fused features, and the Semantic-Aligned Multimodal Reader (SAMR) combines autoregressive generation loss with semantic alignment loss for answer reasoning.

The main contributions of this paper are summarized as follows.

- We propose a multi-source semantic anchor mechanism together with a symmetric cross-modal pre-alignment architecture to bridge the semantic gap among vision, questions, and external knowledge in KB-VQA. This mechanism unifies image captions, object tags, and OCR text as a cross-modal bridge and performs symmetric alignment with visual and question features through VACE and QACE, respectively, transforming complex cross-modal matching into a semantic-to-semantic matching process.
- We design a Cross-Residual Gated Q-V Fusion module (CRGF) to address the modality noise problem in multi-source information fusion. This module treats semantic anchors as a fixed reference and leverages semantic residuals from complementary modalities to drive gating signals that adaptively suppress noise, while Variational Information Bottleneck (VIB) regularization provides complementary assurance from the perspective of representation compression.
- We construct a retrieval–generation collaborative framework jointly optimized by four loss functions. During training, Visual–Anchor Contrastive Learning (VACL) directly constrains the alignment structure of the representation space. During inference, the Dense Knowledge Retriever (DKR) and the Semantic-Aligned Multimodal Reader (SAMR) work in concert to achieve effective complementarity between explicit and implicit knowledge. Experiments on the OK-VQA, FVQA, and A-OKVQA benchmarks validate the effectiveness of the proposed method.

2. Related Work

2.1. Knowledge-Based Visual Question Answering

Knowledge-based visual question answering aims to leverage external knowledge to bridge the cognitive gap between visual content and language questions. According to the manner in which knowledge is acquired, existing methods can be broadly divided into two paradigms: explicit retrieval-based methods and implicit reasoning-based methods.

Explicit retrieval-based methods retrieve relevant evidence from external knowledge bases to enhance reasoning. Early work relies on fixed knowledge bases such as ConceptNet, combined with attention mechanisms for reasoning [3,8,9]. Subsequent studies progressively explore richer knowledge sources, including Wikipedia [10], Google Search [11], and GPT-3 [12]. TRiG [5] constructs textual queries from image captions, tags, and OCR text and retrieves relevant knowledge through dense passage retrieval. RA-VQA [13] performs end-to-end joint training of differentiable dense retrieval and answer generation, and RA-VQA-v2 [14] further improves recall by combining text-based and image-based retrieval. KAT [6] and REVIVE [15] fuse explicit retrieval from Wikipedia with implicit generation from GPT-3 as complementary knowledge sources. Despite steady performance improvements, these methods typically reduce images to a single textual description during the retrieval stage, overlooking fine-grained information such as object attributes and scene text. Moreover, the irrelevant or redundant knowledge introduced by retrieval often produces noise that interferes with downstream reasoning.

Implicit reasoning-based methods treat large language models as parameterized knowledge bases and guide reasoning through prompt engineering. PICa [16] is the first to apply GPT-3 to multimodal tasks, performing reasoning in the language space via few-shot, in-context learning. Prophet [17] refines prompt design using candidate answers, and PromptCap [18] generates question-relevant captions to replace generic descriptions. ASB [19] prompts LLaMA 2 with question-aware captions for training-free reasoning. In addition, end-to-end multimodal large models such as Flamingo [20] and PaLM-E [21] achieve reasoning directly through large-scale vision–language joint pretraining. However, these methods either lose image details due to vision-to-text conversion [22] or require extremely large parameter scales to achieve satisfactory performance.

In contrast to the above methods, the BRIDGE framework simultaneously addresses two commonly overlooked issues. First, it replaces the single caption with multi-source semantic anchors (captions, tags, and OCR text), preserving fine-grained semantic information of the image from multiple complementary perspectives, thereby improving the precision of knowledge retrieval. Second, it explicitly suppresses modality noise introduced during retrieval through a cross-residual gated fusion mechanism, rather than simply concatenating retrieved knowledge with model inputs.

2.2. Cross-Modal Semantic Alignment and Contrastive Learning

Cross-modal semantic alignment aims to narrow the semantic gap between visual and language representations and constitutes a core challenge in VQA tasks. Existing methods primarily rely on attention mechanisms to achieve cross-modal interaction [23,24]. ViLBERT [23] and LXMERT [24] further extend the Transformer architecture to VQA by performing deep cross-modal reasoning through stacked cross-attention layers. However, these methods often overlook the significant semantic gap between different modalities: while unimodal encoders effectively encode visual and textual embeddings within their respective spaces, directly enabling interaction between non-aligned features remains highly challenging [25]. To bridge this gap, Wang et al. [25] propose using image captions as a cross-modal bridge: captions belong to the same textual modality as questions while maintaining close semantic ties with visual content and thus serve as a natural intermediary connecting the two modalities. Their method employs captions as a hub to perform symmetric pre-alignment with visual and question features, respectively, thereby avoiding direct interaction between heterogeneous modalities. However, this approach relies on a single caption as the bridge, which provides insufficient information coverage in KB-VQA scenarios that require external knowledge, and it does not incorporate a noise suppression mechanism during the fusion stage. Beyond the vision–language domain, analogous design principles—such as progressive heterogeneous feature interaction, semantic-guided gating, and adaptive multi-source fusion—have proven effective in other visual tasks, including salient object detection [26], video saliency prediction [27,28], driver attention prediction [29], and change detection [30], suggesting that bridging semantic gaps through intermediate representations and noise-aware fusion constitutes a broadly applicable paradigm.

At the representation learning level, contrastive learning has proven to be an effective means of enhancing cross-modal alignment. CLIP [31] achieves high-quality zero-shot cross-modal representations by jointly training image and text encoders. ALBEF [32] introduces a question–vision (Q-V) contrastive loss in vision–language pretraining to achieve representation pre-alignment before fusion. However, Wang et al. [25] observe that the efficiency of Q-V contrastive learning is limited, because questions provide only a partial and interrogative description of the image, resulting in weak semantic association. Replacing Q-V pairs with vision–caption (V-C) pairs more effectively enhances the cross-modal alignment capability of the encoder. Moreover, since captions and questions share the same text encoder, V-C contrastive learning indirectly improves the perception of visual semantics by the question encoder.

This paper extends the above work in two directions. On the alignment mechanism side, we expand the single-caption bridge to multi-source semantic anchors (captions, tags, and OCR text), bridging the semantic gap between vision and questions from three complementary dimensions: global scene semantics, fine-grained entity attributes, and symbolic text. We further embed a cross-residual gating mechanism in the Q-V fusion stage to continuously guide the fusion process. On the contrastive learning side, we replace the single caption with multi-source semantic anchors to construct the contrastive

learning objective. By exploiting the tighter and more comprehensive semantic association between anchors and visual content, this design more thoroughly constrains the cross-modal alignment structure of the representation space.

3. Proposed Method

3.1. Overall Framework

Existing KB-VQA methods face two core challenges in cross-modal fusion. First, visual features and question/knowledge features reside in heterogeneous modality spaces, and their direct fusion leads to semantic misalignment, making it difficult to precisely localize question-relevant visual content. Second, naive concatenation of multi-source information (visual, question, and external knowledge) introduces modality noise that degrades reasoning accuracy. To address these challenges, we propose the BRIDGE framework, which comprises the following three core modules.

1. Multimodal Feature Extraction Module. This module extracts high-dimensional features from images, questions, and external knowledge. Within it, the Multi-Source Semantic Anchor Construction (MSAC) submodule unifies image captions, object tags, and OCR text into semantic anchors. These anchors share the same textual modality as the question, enabling direct comparison in a unified semantic space; meanwhile, because they originate from the image itself, they inherently encode visual semantics, thereby serving as a bridge that connects visual perception, language understanding, and external knowledge retrieval.
2. Semantic Anchor-Based Cross-Modal Alignment and Fusion Module. Using semantic anchors as intermediaries, this module establishes pre-aligned representations in the visual and question domains through the Visual–Anchor Cross-Modal Encoder (VACE) and the Question–Anchor Cross-Modal Encoder (QACE), respectively, thereby circumventing the semantic misalignment caused by direct interaction between visual and question features. After pre-alignment, the two streams of enhanced features undergo deep fusion in the Cross-Residual Gated Q-V Fusion (CRGF) module, which leverages semantic residuals to adaptively suppress modality noise and employs Variational Information Bottleneck (VIB) regularization to constrain the information efficiency of the fused representation.
3. Cross-Modal Contrastive Learning and Knowledge-Enhanced Reasoning Module. The Visual-Anchor Contrastive Learning (VACL) submodule introduces a cross-modal contrastive learning objective during training, directly constraining the alignment qual-ity between visual features and semantic anchors in a shared space, thereby reinforcing cross-modal semantic consistency at the representation learning level. During inference, the Dense Knowledge Retriever (DKR) performs knowledge retrieval based on the fused features, and the Semantic-Aligned Multimodal Reader (SAMR) generates the final answer guided by a semantic alignment loss.

Cross-Modal Contrastive Learning and Knowledge-Enhanced Reasoning Module. The Visual–Anchor Contrastive Learning (VACL) submodule introduces a cross-modal contrastive learning objective during training, directly constraining the alignment quality between visual features and semantic anchors in a shared space, thereby reinforcing cross-modal semantic consistency at the representation learning level. During inference, the Dense Knowledge Retriever (DKR) performs knowledge retrieval based on the fused features, and the Semantic-Aligned Multimodal Reader (SAMR) generates the final answer guided by a semantic alignment loss.

The overall architecture of the proposed model is illustrated in Figure 2.

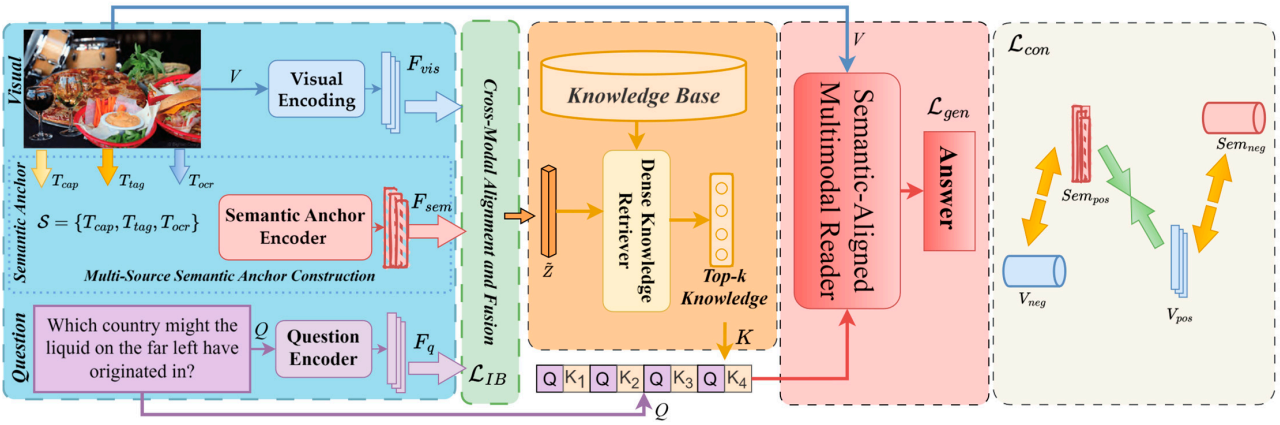


Figure 2. Overall architecture of the BRIDGE framework. The model comprises three stages: (1) Multimodal Feature Extraction, in which a visual encoder extracts image features F_{vis}, a shared RoBERTa encoder independently encodes the question to obtain F_q, and the Multi-Source Semantic Anchor Construction (MSAC) module unifies image captions T_{cap}, object tags T_{tag}, and OCR text T_{ocr} into semantic anchor features F_{sem}; (2) Cross-Modal Alignment and Fusion (detailed in Figure 3), where semantic anchors serve as intermediaries for pre-alignment (VACE/QACE) and cross-residual gated Q-V fusion (CRGF), with Variational Information Bottleneck (VIB) regularization compressing the fused representation $\widetilde{z}$; (3) Knowledge-Enhanced Reasoning, where the Dense Knowledge Retriever (DKR) retrieves top-k relevant entries from the external knowledge base, and the Semantic-Aligned Multimodal Reader (SAMR) generates the final answer supervised by the generation loss $\mathcal{L}_{gen}$. During training, Visual–Anchor Contrastive Learning (VACL) enforces cross-modal alignment via $\mathcal{L}_{con}$ by contrasting positive visual–anchor pairs (V_{pos}, Sem_{pos}) against in-batch negatives.

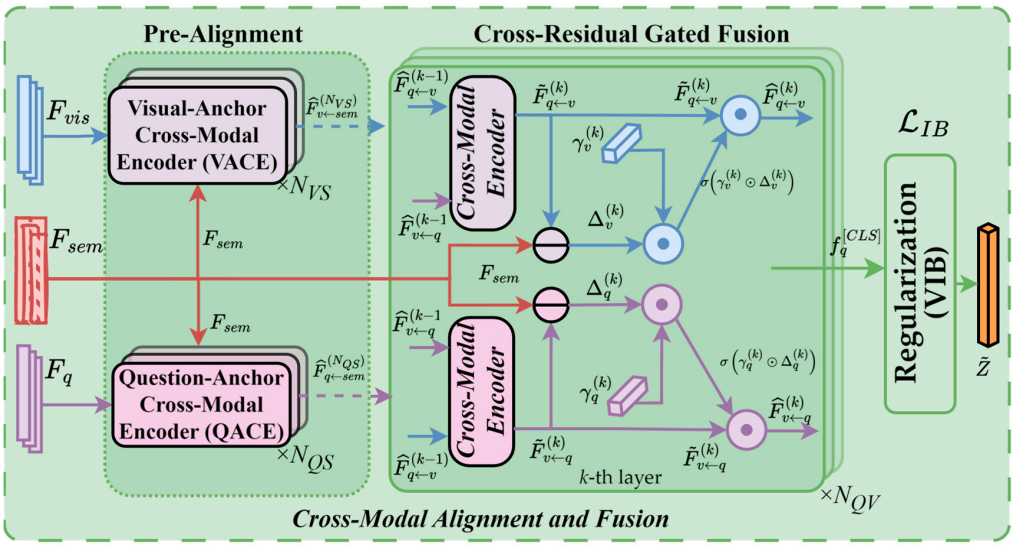


Figure 3. Detailed architecture of the Cross-Modal Alignment and Fusion module. Left (Pre-Alignment stage): The Visual–Anchor Cross-Modal Encoder (VACE) performs N_{VS} layers of symmetric cross-attention between visual features Fvis and semantic anchor features Fsem, yielding anchor-enhanced visual features $\hat{F}_{v\leftarrow sem}^{(N_{VS})}$; symmetrically, the Question–Anchor Cross-Modal Encoder (QACE) stacks N_{QS} layers to produce anchor-enhanced question features $\hat{F}_{q\leftarrow sem}^{(N_{QS})}$. Right (Cross-Residual Gated Fusion stage): At each of the N_{QV} layers, a standard cross-modal encoder block first performs cross-attention interaction between the two pre-aligned feature streams; then the cross-residual gating mechanism computes semantic residuals $\Delta_v^{(k)}$ and $\Delta_q^{(k)}$ between each stream and the fixed semantic anchor reference Fsem, and uses the complementary-modality residual to generate gating signals $\sigma\left(\gamma_v^{(k)} \odot \Delta_v^{(k)}\right)$ and $\sigma\left(\gamma_q^{(k)} \odot \Delta_q^{(k)}\right)$ that adaptively modulate each feature stream. The [CLS] vector $f_q^{[CLS]}$ is extracted from the final-layer output, regularized by VIB ($\mathcal{L}_{IB}$), and used for downstream answer prediction and dense retrieval.

3.2. Multimodal Feature Extraction

To bridge visual perception and language understanding, high-dimensional features are extracted from images, questions, and external knowledge. This section describes the multimodal feature extraction module in detail, which consists of three submodules: visual feature encoding, question and knowledge feature encoding, and Multi-Source Semantic Anchor Construction (MSAC).

3.2.1. Visual Feature Encoding

To preserve fine-grained visual cues in images, we adopt VinVL [33] as the visual encoder without architectural modification to extract the base visual features from the input image. For a given image I, VinVL [33] produces representations at both global and local levels. At the global level, the model aggregates visual features over the entire image with contextual modeling to generate a global visual representation $v_{cls} \in \mathbb{R}^{d_v}$, which captures the overall appearance, scene layout, and global semantic information of the image. At the local level, based on object detection results, the top-K regions of interest (ROIs) with the highest confidence scores are selected, and their region-level visual feature vectors $\{v_1, v_2, \ldots, v_K\}, v_k \in \mathbb{R}^{d_v}$ are extracted to capture key objects and their locally discriminative visual cues. The global visual feature and the region-level visual features are then concatenated to form the base visual feature representation $F_{vis} \in \mathbb{R}^{(K+1)\times d_v}$, where the first position corresponds to the global representation and the remaining positions correspond to the local representations of individual regions.

3.2.2. Question and Knowledge Feature Encoding

For the original question Q and external knowledge entries K, we adopt a shared RoBERTa [34] encoder for feature encoding, following standard practice in vision–language models [23,24]. For structured triplet knowledge entries (e.g., $\langle head, relation, tail \rangle$), we first linearize them into natural language sentences S_K. To facilitate interaction between the question and the knowledge, we concatenate the question text Q and the knowledge sentence S_K and insert special tokens to construct the input sequence:

$$T_{input} = [\text{CLS}] \oplus S_K \oplus [\text{SEP}] \oplus Q \oplus [\text{SEP}], \tag{1}$$

where $\oplus$ denotes sequence concatenation. After feeding T_{input} into a pretrained RoBERTa [34] model, the hidden states of the last layer are extracted as word-level embedding representations $E_w \in \mathbb{R}^{L\times d}$. These features subsequently serve as the initialization for the question features F_q and knowledge features F_k in downstream fusion modules.

3.2.3. Multi-Source Semantic Anchor Construction

Visual features and question features reside in different modality spaces, and their direct interaction faces a severe semantic gap. Inspired by the work of Wang et al. [25], which employs image captions as a cross-modal bridge to mitigate the semantic discrepancy between visual and question representations, we extend this idea to the knowledge-enhanced VQA setting by incorporating not only image captions but also object tags and OCR text to construct Multi-Source Semantic Anchors. Compared with captions alone, multi-source semantic anchors characterize image content from three complementary dimensions—global scene semantics (captions), fine-grained entity attributes (tags), and symbolic information (OCR)—thereby providing more comprehensive coverage of visual semantics.

Specifically, the MSAC module constructs an explicit semantic set $\mathcal{S} = \{T_{cap}, T_{tag}, T_{ocr}\}$ for image I, where T_{cap}, T_{tag}, and T_{ocr} denote the image caption, object tag text, and OCR text generated from I, respectively. For the image caption T_{cap}, we employ the Oscar [35] model to generate descriptive sentences that capture the global events and scene context

of the image. For the object tags, we apply joint object detection and attribute recognition to model key entities and their attributes [33], yielding a textual representation $T_{tag} = \left\{ (o_j, a_j) \right\}_{j=1}^{M}$, where o and a denote the object and its attribute, respectively. Furthermore, to compensate for the limited capability of visual models in modeling symbolic and textual semantics, we adopt the open-source Umi-OCR [36] model to recognize scene text in the image and extract OCR text T_{ocr} to enhance the perception of fine-grained symbolic semantics.

The text sequences T_{cap}, T_{tag}, and T_{ocr} are then concatenated and encoded through a shared RoBERTa [34] encoder to obtain the semantic anchor features $F_{sem} \in \mathbb{R}^{N_s \times d}$. It is important to note that, during encoding, we strictly maintain the independence between the semantic anchors and the question: although both share the same RoBERTa [34] encoder parameters, they are encoded independently during the forward pass without any cross-interaction. This design follows the paradigm in [25], where the caption encoder and the question encoder perform independent encoding before cross-modal alignment, ensuring that each modality retains its modality-specific semantics before entering the subsequent cross-modal alignment module and preventing premature information mixing from causing the loss of discriminative information. The cross-modal interaction between the semantic anchors and the question features is deferred to the semantic anchor-based cross-modal alignment module (Section 3.3), where it is accomplished through dedicated cross-attention mechanisms. Beyond the pre-alignment role served by captions in [25], the multi-source semantic anchors in our framework simultaneously function as reference targets for residual-gated fusion (Section 3.3.2), anchoring samples for contrastive learning (Section 3.6), and semantic enhancers for knowledge retrieval (Section 3.4), considerably broadening their utility within the pipeline.

3.3. Cross-Modal Alignment and Fusion

The semantic gap between different modalities hinders effective interaction and alignment between visual features and question features. To address this problem, we design a two-stage cross-modal alignment and fusion module with semantic anchors as intermediaries, following a "pre-alignment followed by deep fusion" paradigm. In the pre-alignment stage, semantic anchors interact symmetrically with visual features and question features through the Visual–Anchor Cross-Modal Encoder (VACE) and the Question–Anchor Cross-Modal Encoder (QACE), respectively, establishing preliminary aligned representations for each modality under the bridging of semantic anchors. In the deep fusion stage, the two streams of pre-aligned enhanced features enter the Cross-Residual Gated Q-V Fusion (CRGF) module for deep fusion, which incorporates a cross-residual gating modulation mechanism to suppress modality noise.

All submodules described above are built upon a unified cross-modal encoder block, which follows the cross-modal interaction architecture widely used in vision–language models such as ViLBERT [23] and LXMERT [24]. Compared with a unimodal encoder block that contains only self-attention (SA) and a feed-forward network (FFN), the cross-modal encoder block prepends an additional group of cross-attention (CA), residual connection, and layer normalization operations, enabling cross-modal information exchange between two feature streams. Specifically, this encoder block takes two feature streams as input, and its symmetric architecture allows both streams to simultaneously inject information into each other. Taking one stream as an example, at the k-th layer, this stream serves as the query while the other stream serves as both the key and the value for the cross-attention operation. The output then undergoes self-attention and the feed-forward network for internal refinement, producing an enhanced representation that incorporates information from the counterpart stream. In the pre-alignment stage, VACE and QACE directly adopt

this base block for semantic bridging, whereas CRGF in the deep fusion stage further introduces a cross-residual gating modulation mechanism on top of this block to suppress modality noise. The detailed architecture of the cross-modal alignment and fusion module is illustrated in Figure 3.

3.3.1. Pre-Alignment

To avoid the semantic misalignment caused by direct interaction between visual features and question features, we employ semantic anchors F_{sem} as intermediaries to establish pre-aligned representations in the visual domain and the question domain, respectively. This symmetric pre-alignment design is motivated by the observation that, in KB-VQA, the semantic gap spans three dimensions—vision, questions, and external knowledge—rather than the two-dimensional vision–question gap addressed in standard VQA. A single caption, as used in [25], cannot adequately anchor all three dimensions: it lacks fine-grained entity attributes and symbolic information critical for knowledge retrieval. Our design therefore introduces multi-source semantic anchors as the intermediary and decouples pre-alignment from deep fusion into a two-stage paradigm, enabling the anchors to first establish modality-specific alignment before cross-modal noise is introduced in the fusion stage.

(1) Visual–Anchor Cross-Modal Encoder

VACE stacks N_{VS} cross-modal encoder blocks, enabling symmetric cross-attention interaction between semantic anchors and visual features. With F_{sem} and the linearly projected visual features $W_v F_{vis}$ as the initial inputs, the computation at the k-th layer proceeds as follows:

$$\hat{F}^{(k)}_{sem\leftarrow v} = \mathrm{FFN}\Big(\mathrm{SA}\Big(\mathrm{CA}\Big(\hat{F}^{(k-1)}_{sem\leftarrow v}, \hat{F}^{(k-1)}_{v\leftarrow sem}\Big)\Big)\Big), \tag{2}$$

$$\hat{F}^{(k)}_{v\leftarrow sem} = \mathrm{FFN}\Big(\mathrm{SA}\Big(\mathrm{CA}\Big(\hat{F}^{(k-1)}_{v\leftarrow sem}, \hat{F}^{(k-1)}_{sem\leftarrow v}\Big)\Big)\Big), \tag{3}$$

where $\hat{F}^{(0)}_{sem\leftarrow v} = F_{sem}$, $\hat{F}^{(0)}_{v\leftarrow sem} = W_v F_{vis}$, $W_v \in \mathbb{R}^{d_v \times d}$ is a linear projection matrix that maps the visual dimension to the semantic dimension. $\mathrm{CA}(\cdot,\cdot)$ denotes the cross-attention operation, where the first argument serves as the query and the second serves as both the key and the value. $\mathrm{SA}(\cdot)$ and $\mathrm{FFN}(\cdot)$ denote self-attention and feed-forward network operations, respectively; the subsequent residual connections and layer normalization are omitted for brevity.

After N_{VS} layers of alignment, the output of the final layer $\hat{F}^{(N_{VS})}_{v\leftarrow sem}$ constitutes the anchor-enhanced visual representation, in which each visual token has aggregated its corresponding semantic descriptive information.

(2) Question–Anchor Cross-Modal Encoder

Similarly, QACE stacks N_{QS} cross-modal encoder blocks, enabling symmetric cross-attention interaction between semantic anchors and question features:

$$\hat{F}^{(k)}_{sem\leftarrow q} = \mathrm{FFN}\Big(\mathrm{SA}\Big(\mathrm{CA}\Big(\hat{F}^{(k-1)}_{sem\leftarrow q}, \hat{F}^{(k-1)}_{q\leftarrow sem}\Big)\Big)\Big), \tag{4}$$

$$\hat{F}^{(k)}_{q\leftarrow sem} = \mathrm{FFN}\Big(\mathrm{SA}\Big(\mathrm{CA}\Big(\hat{F}^{(k-1)}_{q\leftarrow sem}, \hat{F}^{(k-1)}_{sem\leftarrow q}\Big)\Big)\Big), \tag{5}$$

where $\hat{F}^{(0)}_{sem\leftarrow q} = F_{sem}$, $\hat{F}^{(0)}_{q\leftarrow sem} = F_q$. The output of the final layer $\hat{F}^{(N_{QS})}_{q\leftarrow sem}$ constitutes the anchor-enhanced question representation, in which the originally static question features have been endowed with semantic context related to the visual content.

Through this pre-alignment process, semantic anchors serve as a bridge that injects visual semantics into the question representation and question intent into the visual representation, respectively, so that both feature streams possess a preliminary cross-modal semantic alignment basis before entering the subsequent Q-V fusion.

3.3.2. Cross-Residual Gated Fusion

The anchor-enhanced visual features $\hat{F}_{v\leftarrow sem}^{(N_{VS})}$ and anchor-enhanced question features $\hat{F}_{q\leftarrow sem}^{(N_{QS})}$ produced by VACE and QACE in the pre-alignment stage subsequently enter CRGF for deep fusion. In the pre-alignment stage, the standard cross-modal encoder block is sufficient for establishing preliminary semantic bridging. However, the Q-V fusion stage faces a more challenging modality noise problem: although the enhanced visual and question features have been preliminarily aligned, they may still retain visual redundancies irrelevant to the current question or question components that do not match the image content. To address this, CRGF introduces a Cross-Residual Gating mechanism on top of the standard cross-modal encoder block. The key idea is to use the semantic anchor F_{sem} as a fixed reference that represents the target semantics of the image. At each fusion layer, the semantic residual between F_{sem} and each feature stream quantifies how far the current representation deviates from the target; the residual from the complementary modality then drives a gating signal that modulates the current modality, achieving dynamic inter-modal complementarity.

Specifically, CRGF stacks N_{QV} layers in total. At the k-th layer, feature interaction is first performed through the standard cross-modal encoder block:

$$\tilde{F}_{q\leftarrow v}^{(k)} = \mathrm{FFN}\Big(\mathrm{SA}\Big(\mathrm{CA}\Big(\hat{F}_{q\leftarrow v}^{(k-1)}, \hat{F}_{v\leftarrow q}^{(k-1)}\Big)\Big)\Big), \tag{6}$$

$$\tilde{F}_{v\leftarrow q}^{(k)} = \mathrm{FFN}\Big(\mathrm{SA}\Big(\mathrm{CA}\Big(\hat{F}_{v\leftarrow q}^{(k-1)}, \hat{F}_{q\leftarrow v}^{(k-1)}\Big)\Big)\Big), \tag{7}$$

where the initial inputs are $\hat{F}_{q\leftarrow v}^{(0)} = \hat{F}_{q\leftarrow sem}^{(N_{QS})}$ and $\hat{F}_{v\leftarrow q}^{(0)} = \hat{F}_{v\leftarrow sem}^{(N_{VS})}$.

Cross-residual gating modulation is then applied to the output of each layer. Using the semantic anchor F_{sem} as a fixed reference, the semantic residual between each of the two feature streams and the anchor at the current layer is computed:

$$\Delta_q^{(k)} = F_{sem} - \tilde{F}_{q\leftarrow v}^{(k)}, \quad \Delta_v^{(k)} = F_{sem} - \tilde{F}_{v\leftarrow q}^{(k)}, \tag{8}$$

The cross-residual gating operation is then performed, which uses the semantic residual of the complementary modality as a gating signal to modulate the features of the current modality:

$$\hat{F}_{q\leftarrow v}^{(k)} = \tilde{F}_{q\leftarrow v}^{(k)} \odot \sigma\Big(\gamma_v^{(k)} \odot \Delta_v^{(k)}\Big), \tag{9}$$

$$\hat{F}_{v\leftarrow q}^{(k)} = \tilde{F}_{v\leftarrow q}^{(k)} \odot \sigma\Big(\gamma_q^{(k)} \odot \Delta_q^{(k)}\Big), \tag{10}$$

where $\odot$ denotes element-wise multiplication, $\sigma(\cdot)$ is the Sigmoid activation function, and $\gamma_v^{(k)}$ and $\gamma_q^{(k)}$ are independent learnable scaling parameters at the k-th layer. The physical interpretation of the above formulas is as follows: when the semantic residual $\Delta_{\overline{m}}^{(k)}$ of the complementary modality is large, it indicates that this modality cannot sufficiently explain the image content described by the semantic anchors at the current layer, and the model forces the current modality to assume greater information transfer responsibility through the gating mechanism, thereby achieving dynamic complementarity between the two modalities.

We choose a semantic anchor F_{sem} as a fixed reference target shared across all layers, rather than using the incremental change between adjacent layers as the residual basis. This design ensures that the gating mechanism always operates based on the absolute deviation between the current representation and the target semantics. If the residual were computed relative to the output of the previous layer, the differences between adjacent layers would tend to become smooth as depth increases due to the diminishing-increment effect in deep networks, causing the gating signal to attenuate progressively across layers. By contrast, the fixed-anchor scheme ensures that the gating signal remains active even in deeper layers as long as the fused features have not yet adequately expressed the anchor semantics, thereby maintaining the refinement efficacy across all layers.

Compared with a single-step gating iteration that involves only element-wise scaling, embedding the cross-residual gating mechanism inside the cross-modal encoder block offers two advantages. First, the CA $\rightarrow$ SA $\rightarrow$ FFN pipeline provides sufficient feature recombination and nonlinear transformation capacity, enabling each layer to perform reasoning at a higher level of semantic abstraction. Second, the gating operation is applied to the FFN output—at which point the features have already undergone deep interaction through cross-attention and self-attention—making the semantic residual signal more discriminative and the gating decisions consequently more precise.

3.3.3. Answer Representation Extraction and Regularization

From the output of the final CRGF layer $\hat{F}_{q\leftarrow v}^{(N_{QV})}$, the vector at the [CLS] position $f_q^{[\mathrm{CLS}]} \in \mathbb{R}^d$ is extracted as the answer representation. This representation aggregates cross-modal semantic information from both the pre-alignment and deep fusion stages and serves as the shared basis for subsequent answer prediction and knowledge retrieval. However, although the cross-residual gating mechanism in CRGF suppresses modality noise at the feature interaction level, it is inherently deterministic—for a given combination of input features, the gating output is fixed, and the fused representation may still retain redundant details irrelevant to the downstream task. To this end, we introduce Variational Information Bottleneck (VIB) regularization [7], which formalizes the information bottleneck principle within a deep learning framework, providing complementary assurance from the perspective of representation compression, forcing the model to encode only the semantics most critical for answer prediction within a limited information capacity. Specifically, $f_q^{[\mathrm{CLS}]}$ is treated as the parameterization of a conditional distribution $q_\theta\left(z \mid f_q^{[\mathrm{CLS}]}\right) = \mathcal{N}\left(\mu, \mathrm{diag}(\sigma^2)\right)$:

$$\mu = f_q^{[CLS]}, \quad \log\sigma^2 = W_\sigma f_q^{[CLS]} + b_\sigma, \tag{11}$$

where $W_\sigma \in \mathbb{R}^{d\times d}$ and $b_\sigma \in \mathbb{R}^d$ are learnable parameters. During training, a stochastic representation $\widetilde{z} = \mu + \sigma \odot \epsilon$ (where $\epsilon \sim \mathcal{N}(0, I)$) is sampled from this distribution via the reparameterization trick and used in place of $f_q^{[\mathrm{CLS}]}$ for subsequent classification and dense retrieval; during inference, the mean μ is directly used as the deterministic output. The corresponding VIB loss is defined as the Kullback–Leibler (KL) divergence between this conditional distribution and the standard normal prior:

$$\mathcal{L}_{IB} = \frac{1}{2d}\sum_{j=1}^{d}\left(\mu_j^2 + \sigma_j^2 - \ln\sigma_j^2 - 1\right), \tag{12}$$

$\mathcal{L}_{IB}$ drives the fused representation toward a compact prior distribution, compressing modality noise and redundant details while retaining the core semantic information relevant to answer prediction.

After the above regularization, the answer representation is utilized along two pathways. The first pathway feeds it into a two-layer Multi-Layer Perceptron (MLP) classifier to generate a probability distribution over candidate answers:

$$\hat{y} = \text{softmax}\Big(W_2 \cdot \text{ReLU}\Big(W_1 \cdot f_q^{[CLS]}\Big)\Big), \tag{13}$$

where $\hat{y} \in \mathbb{R}^{A_c}$ is the predicted distribution, and A_c is the number of candidate answers. Concurrently, $f_q^{[\text{CLS}]}$ is linearly projected to obtain the query vector for the Dense Knowledge Retriever (DKR): $\overline{F}_{query} = W_{proj} f_q^{[\text{CLS}]}$, which is used for subsequent knowledge base retrieval.

3.4. Dense Knowledge Retriever

Open-domain visual question answering requires external knowledge to bridge the semantic gap between visual content and linguistic questions. To this end, the Dense Knowledge Retriever (DKR) adopts the dense passage retrieval paradigm [37], which maps queries and knowledge entries into a unified continuous vector space, enabling efficient retrieval based on semantic similarity. This paradigm has been successfully applied to KB-VQA by methods such as TRiG [5] and RA-VQA [13]. Our DKR follows this established framework but differs in the query construction: rather than using a text-only query, our query vector is derived from the cross-modal fused representation (Section 3.3.3), allowing the retrieval to leverage the multi-modal alignment achieved by the upstream modules. Compared with sparse methods such as BM25 or TF-IDF, this approach more effectively captures the deep associations between cross-modal features and textual knowledge.

Specifically, DKR takes the query vector $\overline{F}_{query}$, obtained by linearly projecting the answer representation $f_q^{[\text{CLS}]}$ from Section 3.3.3 (replaced by the stochastic representation $\widetilde{z}$ after VIB randomization during training), as the query representation for the current instance to drive subsequent retrieval. To support reasoning, we define an external knowledge base $\mathcal{K} = \{K_1, K_2, \ldots, K_M\}$. For the i-th knowledge entry K_i in the knowledge base, a pre-trained independent BERT encoder maps it into the same semantic space, yielding a knowledge embedding vector E_{K_i}:

$$E_{K_i} = f_{enc}(K_i; \theta_K), \tag{14}$$

where $f_{enc}(\cdot)$ denotes the independent BERT-architecture encoder and θ_K denotes its trainable parameters.

In the unified vector space, cosine similarity is used to measure the semantic relevance between the query $\overline{F}_{query}$ and the knowledge embedding E_{K_i}:

$$\text{Sim}\big(\overline{F}_{query}, K_i\big) = \frac{\overline{F}_{query}^{\top} E_{K_i}}{\| \overline{F}_{query} \|_2 \cdot \| E_{K_i} \|_2}, \tag{15}$$

The knowledge base is ranked by similarity scores, and the top-k entries with the highest scores are selected to form the candidate knowledge subset: $\mathcal{K}_{ret} = \{K_{p_1}, K_{p_2}, \ldots, K_{p_k}\}$. This subset, representing the most relevant external knowledge, serves as explicit factual evidence in the subsequent reasoning stage, enhancing both the reasoning capability and the interpretability of the model.

3.5. Semantic-Aligned Multimodal Reader

After retrieving external knowledge, a reader with multimodal understanding and text generation capabilities is required to deeply comprehend the candidate knowledge subset $\mathcal{K}_{ret} = \{K_{p_1}, K_{p_2}, \ldots, K_{p_k}\}$ and generate the final answer. Unlike discriminative

models such as ViLT [38], LXMERT [24], and VisualBERT [39], which formulate VQA as a classification problem over a fixed candidate set, generative multimodal models treat VQA as an open-ended sequence generation task, exhibiting stronger flexibility and generalization capability. Accordingly, we adopt a generative multimodal model as the reader backbone and introduce a semantic alignment enhancement mechanism on top of it, constructing the Semantic-Aligned Multimodal Reader (SAMR). The design of this module is decoupled from any specific generative backbone and can be adapted to various mainstream multimodal generative models (e.g., BLIP [40], Qwen2-VL [41], LLaVA-NeXT [42]); a systematic comparison of different backbones is provided in the experimental section.

Specifically, given the raw image I, the question Q, and a single candidate knowledge entry K_{p_i}, the reader backbone obtains a context representation that fuses multimodal information through its built-in visual encoder and cross-modal interaction mechanism and then autoregressively generates the answer via the decoder. To adapt to the KB-VQA task, SAMR introduces a semantic alignment loss on top of the standard autoregressive generation loss. The generation loss $\mathcal{L}_{gen}$ consists of two complementary objectives.

The first objective is the standard autoregressive generation loss, which constrains the decoder at the token level to sequentially generate a word sequence consistent with the ground-truth answer:

$$\mathcal{L}_{AR} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i}\log P_\theta\left(a_t^i \middle| a_{<t}^i, I_i, Q_i, K_{p_i}\right), \tag{16}$$

where a_t^i is the t-th token of the ground-truth answer for the i-th sample, T_i is the answer length, and P_θ is the conditional probability of the decoder. $\mathcal{L}_{AR}$ ensures the grammatical fluency and surface-level correctness of the generated output.

However, in VQA tasks, the same question often admits multiple semantically equivalent correct expressions (e.g., "dog" and "puppy", "2" and "two"), and purely token-level matching would incorrectly penalize synonymous answers. To address this, rather than modifying the token-level loss function, SAMR introduces a second objective, the semantic alignment loss $\mathcal{L}_{SA}$, which operates at the sentence-level semantic space to constrain overall generation quality. Specifically, a frozen pre-trained BERT is used to encode the [CLS] vectors of the generated answer G and the ground-truth answer A:

$$h_G = \mathrm{BERT}_{[CLS]}(G), \quad h_A = \mathrm{BERT}_{[CLS]}(A), \tag{17}$$

The two vectors are then concatenated and passed through a linear mapping followed by Softmax to compute the semantic matching probability:

$$P_{match} = \mathrm{softmax}(W_s[h_G; h_A] + b_s), \tag{18}$$

where $W_s \in \mathbb{R}^{2\times 2d}$ and $b_s \in \mathbb{R}^2$ are learnable parameters, $[\cdot;\cdot]$ denotes vector concatenation, and $P_{match} \in \mathbb{R}^2$ is the binary classification probability (semantic match/non-match). The semantic alignment loss is defined as

$$\mathcal{L}_{SA} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log P_{match,1}^i + (1-y_i)\log P_{match,0}^i\right], \tag{19}$$

where $y_i = 1$ indicates that the generated answer semantically matches the ground-truth answer. The total generation-stage loss of SAMR is defined as

$$\mathcal{L}_{gen} = \mathcal{L}_{AR} + \lambda\mathcal{L}_{SA}, \tag{20}$$

where λ is a balancing coefficient. $\mathcal{L}_{AR}$ constrains the step-by-step correctness of generation at the token level, while $\mathcal{L}_{SA}$ constrains the overall semantic consistency of generation at the semantic level; together, they supervise the quality of answer generation from complementary perspectives.

3.6. Visual–Anchor Contrastive Learning

In addition to the generation-stage losses described above, the Visual–Anchor Contrastive Learning (VACL) module introduces a cross-modal contrastive learning objective at the representation learning stage, directly acting on the geometric structure of the representation space to reinforce the alignment quality between visual features and semantic anchors in the shared semantic space.

The effectiveness of cross-modal contrastive learning depends critically on the semantic tightness between paired samples. Question–visual (Q-V) pairs suffer from weak alignment because questions only partially and interrogatively refer to the image, capturing neither its global scene nor its fine-grained details. Since the multi-source semantic anchors constructed in Section 3.2.3 directly describe image content from three complementary dimensions—global scene semantics, entity attributes, and symbolic information—they naturally exhibit a tighter semantic association with visual content than questions do. We therefore construct visual–anchor (V-Sem) pairs as contrastive training samples to more thoroughly constrain the cross-modal alignment structure of the representation space. Moreover, since the semantic anchors and the question share the same RoBERTa [34] encoder, optimizing the cross-modal alignment capability of the anchor encoder through VACL implicitly improves the question encoder's perception of visual semantics, thereby achieving implicit enhancement of question–visual alignment without introducing additional parameters. A systematic comparison of V-Sem pairs against alternative pairing strategies, including Q-V and caption–visual (V-C) [25] pairs, is provided in the ablation study (Section 4.6.3).

3.6.1. Global Feature Extraction

For the i-th sample in the batch, the global representation v_{cls}^{i} from the visual encoder (the global visual representation defined in Section 3.2.1) and the vector at the [CLS] position $f_{sem}^{[\mathrm{CLS}],i}$ from the semantic anchor encoder output are extracted and then projected into a low-dimensional shared space through their respective projection heads:

$$z_{vis}^{i} = \mathrm{L2Norm}\left(\mathrm{MLP}_{vis}\left(v_{cls}^{i}\right)\right), \tag{21}$$

$$z_{sem}^{i} = \mathrm{L2Norm}\left(\mathrm{MLP}_{sem}\left(f_{sem}^{[CLS],i}\right)\right), \tag{22}$$

where MLP_{vis} and MLP_{sem} are two-layer MLP projection heads with an output dimension of d_p.

3.6.2. Negative Sample Construction

Within each training batch, the semantic anchor global representation z_{sem}^{i} of the i-th sample serves as the anchor sample, and the visual global representation z_{vis}^{i} of the corresponding image constitutes the positive sample. The visual representations of the remaining $B-1$ samples $\left\{z_{vis}^{j}\right\}_{j\neq i}$ in the batch form the negative sample set for this anchor; symmetrically, when z_{vis}^{i} serves as the anchor, the semantic anchor representations of the remaining $B-1$ samples $\left\{z_{sem}^{j}\right\}_{j\neq i}$ form its negative sample set. Since the same image may correspond to multiple different question–anchor instances, a binary mask matrix $M \in \{0,1\}^{B\times B}$ is introduced into the similarity matrix to prevent samples originating from

the same image from being incorrectly treated as negative samples, where $M_{ij} = 0$ if and only if samples i and j originate from the same image, in which case the sample pair is excluded from the contrastive loss computation.

3.6.3. Symmetric InfoNCE Loss

VACL adopts a symmetric InfoNCE loss [43], following the cross-modal contrastive learning paradigm established by CLIP [31] and ALBEF [32], to measure the degree of cross-modal alignment:

$$\mathcal{L}_{s2v} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\text{sim}\left(z_{sem}^{i}, z_{vis}^{i}\right)/\tau\right)}{\sum_{j=1}^{B} M_{ij} \cdot \exp\left(\text{sim}\left(z_{sem}^{i}, z_{vis}^{j}\right)/\tau\right)}, \tag{23}$$

$$\mathcal{L}_{v2s} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\text{sim}\left(z_{vis}^{i}, z_{sem}^{i}\right)/\tau\right)}{\sum_{j=1}^{B} M_{ij} \cdot \exp\left(\text{sim}\left(z_{vis}^{i}, z_{sem}^{j}\right)/\tau\right)}, \tag{24}$$

$$\mathcal{L}_{con} = \frac{1}{2}(\mathcal{L}_{s2v} + \mathcal{L}_{v2s}), \tag{25}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, τ is a temperature hyperparameter, and B is the batch size. $\mathcal{L}_{con}$ drives the visual global features and the semantic anchor global features of the same image to be close in the projection space while pushing apart the features of other samples within the batch, thereby directly optimizing the alignment structure of the cross-modal representation space during training.

3.7. Total Loss Function

Combining the loss functions from the stages described above, the total training objective of the BRIDGE framework is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{gen} + \alpha\mathcal{L}_{con} + \beta\mathcal{L}_{IB} = \mathcal{L}_{AR} + \lambda\mathcal{L}_{SA} + \alpha\mathcal{L}_{con} + \beta\mathcal{L}_{IB}, \tag{26}$$

where α, β and λ are balancing coefficients. The four loss terms operate at different levels: $\mathcal{L}_{AR}$ constrains the step-by-step generation quality of answers at the output end; $\mathcal{L}_{SA}$ constrains the overall semantic consistency between generated answers and ground-truth answers at the output end; $\mathcal{L}_{con}$ constrains the cross-modal alignment structure of the representation space, ensuring that visual features and semantic anchors are semantically aligned in the shared space; and $\mathcal{L}_{IB}$ constrains the information efficiency of the fused representation by compressing redundant information irrelevant to the task. This multi-level joint optimization strategy ensures the coordinated improvement of the model across representation generation, feature fusion, and answer reasoning.

4. Experiments

4.1. Datasets and Knowledge Bases

To comprehensively evaluate the performance of the BRIDGE framework, this paper conducts experiments on three mainstream KB-VQA benchmark datasets.

OK-VQA [44] is one of the most challenging datasets in this field. Its distinguishing characteristic is that answering the questions heavily depends on external knowledge beyond the image, and no explicit background context is provided. The dataset covers 11 diverse categories, including vehicles, sports, and cooking, and contains approximately 14,055 questions, imposing demanding requirements on the model's generalized knowledge reasoning capability.

FVQA [2] is the first dataset to provide explicit supporting facts. Each sample consists of an image, a question, an answer, and a corresponding supporting knowledge entry,

totaling 5286 questions. It is primarily used to evaluate the reasoning accuracy of models given explicit structured knowledge.

A-OKVQA [45], as an enhanced version of OK-VQA [44], contains approximately 25,000 questions and requires models to reason with more diverse types of external knowledge, encompassing multiple capability dimensions such as commonsense reasoning, world knowledge, and visual knowledge. The dataset provides both Multiple-Choice (MC) and Direct Answer (DA) evaluation settings and annotates reasoning rationales for each question in the training set. It has become a more prevalent evaluation benchmark in the KB-VQA field in recent years.

To support multi-modal reasoning, this study employs two publicly available, static knowledge corpora that are merged into a unified knowledge base K for dense retrieval. The first is a Wikipedia corpus, specifically the OKVQA_passages subset (344K entries) from the M2KR benchmark, which is the same static Wikipedia corpus used by TRiG [5], RA-VQA [13], and RA-VQA-v2 [14]. It provides broad and detailed general world knowledge and serves as the primary source from which the Dense Knowledge Retriever (DKR) obtains open-domain explicit knowledge. The second is a Google Search corpus (okvqa_full_corpus, 168,306 entries) pre-collected by Luo et al. [46] based on the OK-VQA training and testing data. This corpus supplements Wikipedia by covering web-sourced information that may not appear in encyclopedia-style entries, thereby improving knowledge coverage on diverse and long-tail queries. Importantly, both corpora are pre-collected static datasets rather than real-time retrieval results; no live API calls to Wikipedia or Google Search are made during training or inference, ensuring full reproducibility across experimental runs.

4.2. Evaluation Metrics

Following the standard evaluation protocol in the VQA field, this paper adopts VQA accuracy as the core metric on the OK-VQA [44] and A-OKVQA [45] datasets. This metric accommodates the diversity of answers in open-domain question answering through a soft voting mechanism, comparing the model-generated answer a against independent reference answers provided by ten human annotators:

$$\mathrm{Acc}(a) = \min\left(1, \frac{\text{humans that provided } a}{3}\right), \tag{27}$$

A prediction is considered fully correct if the generated answer matches at least three reference answers. For the FVQA [2] dataset, this paper adopts Top-1 and Top-3 accuracy to evaluate model performance on structured knowledge reasoning, where Top-1 measures the absolute precision of the highest-confidence prediction, and Top-3 assesses the recall capability of the knowledge retrieval and ranking stage.

4.3. Compared Methods

To comprehensively evaluate the performance of the BRIDGE framework, this paper categorizes the compared methods into three groups according to their technical approaches.

4.3.1. Retrieval-Based Methods

Retrieval-based methods typically adopt a retriever-reader framework that retrieves knowledge relevant to the question and image from external knowledge bases and then combines visual and textual information for reasoning to generate answers.

ConceptBERT [8] uses ViLBERT [23] to learn fused features of questions and images, extracts commonsense knowledge from ConceptNet, encodes knowledge triplets through a

graph convolutional network, and jointly learns the interactions among images, questions, and knowledge.

KRISP [47] combines implicit knowledge with explicit symbolic knowledge. The implicit reasoning component uses Multi-modal BERT for question and image encoding, while the explicit reasoning component queries relevant knowledge nodes based on a graph network. It fuses four knowledge sources: DBPedia, ConceptNet, VisualGenome, and hasPart KB.

RVL [9] incorporates knowledge vectors into the training stage of LXMERT [24], using ConceptNet and Wikipedia as explicit knowledge sources and enhancing reasoning through the alignment of knowledge graph embeddings with cross-modal representations.

VRR [46] employs a separate retriever to retrieve knowledge from ConceptNet and Google Search and trains the retriever using retrieved passages containing the answer as weak supervision signals.

MAVEx [48] performs answer-guided knowledge retrieval. Beyond textual knowledge from Wikipedia and ConceptNet, it innovatively uses the Google image search engine to retrieve external visual knowledge and conducts multi-modal answer verification across different knowledge sources.

UnifER+ViLT [3] proposes a unified end-to-end retriever-reader framework. The reader component is based on ViLT [38] and fuses question-image features, retrieving explicit knowledge from ConceptNet while incorporating implicit knowledge from vision–language pretrained models.

KAT-T5 [6] proposes a knowledge-augmented Transformer that integrates implicit and explicit knowledge within an encoder–decoder architecture, using a contrastive learning module to retrieve explicit knowledge from Wikipedia while leveraging the implicit knowledge of T5-large.

LaKo+T5 [49] uniformly converts structured knowledge from ConceptNet, DBPedia, and WebChild into the textual modality to fully exploit the semantic understanding capability of the T5 model.

TriG [5] uses a dense knowledge retriever to extract relevant knowledge passages from Wikipedia, converts images into textual features, and then uses T5-large to integrate multi-source information and generate answers.

RA-VQA [13] proposes a retrieval-augmented visual question answering framework that performs end-to-end joint training of differentiable dense passage retrieval and answer generation, enabling the retriever and generator to mutually optimize each other.

RA-VQA-v2 [14], as its improved version, proposes a fine-grained late-interaction multi-modal retrieval method, uses T5-XL (3B parameters) as the generator, and improves retrieval recall by incorporating visual features to supplement textual descriptions.

4.3.2. LLM Prompt-Based Methods

Large Language Model (LLM) prompt-based methods treat large language models as implicit knowledge bases and guide knowledge acquisition and reasoning through prompt engineering.

KAT-base [6] jointly leverages explicit and implicit knowledge. It designs textual prompts based on image captions and uses GPT-3 (175B) to retrieve implicit knowledge, achieving complementary fusion of the two knowledge sources.

DEDR+MM-FiD [4] retrieves knowledge from Wikipedia through both unimodal and multimodal symmetric dense retrieval schemes, combines GPT-3 and LXMERT models, and fuses multiple retrieved documents using a Fusion-in-Decoder strategy.

PICa-Full [16] is the first method to apply GPT-3 to multimodal tasks. It converts images into textual captions and tags and then guides GPT-3 reasoning through few-shot in-context learning.

KAT (Single/Ensemble) [6] replaces T5 with GPT-3 as the implicit knowledge source on the basis of KAT-T5 [6], and the Ensemble version further improves performance through multi-model ensembling.

REVIVE [15] improves knowledge retrieval by leveraging regional visual features, integrating visual information, explicit knowledge from Wikipedia, and implicit knowledge from GPT-3.

PromptCap [18] proposes a prompt-guided image captioning method that generates question-relevant captions and uses GPT-3 for training sample synthesis and answer reasoning.

MM-Reasoner [50] combines GPT-4 and i-Code v2 models, integrating visual understanding and external knowledge through a multi-stage reasoning pipeline.

Prophet [17] extracts candidate answers and answer-aware examples from a pretrained VQA model to enhance the prompts for GPT-3, using GPT-3 only at inference time.

ASB [19] uses question-aware captions to prompt LLaMA 2 (13B) through efficient in-context learning, requiring no training or access to external databases.

SKP [51] proposes a soft knowledge prompting method that extracts valuable information from external knowledge to generate latent vectors as soft prompts, which are fused with image embeddings to guide Vicuna-7b reasoning.

PrQAC [52] generates question-aware captions and candidate answers to construct structured prompts that guide LLaMA-3 (8B) in answer generation.

GC-KBVQA [53] proposes a grounded captioning-based four-stage framework that enables LLaMA-3 to perform VQA tasks in a zero-shot setting.

4.3.3. End-to-End Multi-Modal Large Models

With the rapid development of end-to-end multi-modal large models, the following methods achieve multi-modal reasoning directly through large-scale vision–language joint pretraining, without requiring explicit knowledge retrieval or prompt engineering steps. Instead, they take images and questions as input and perform end-to-end reasoning.

Flamingo [20] (80B) achieves strong few-shot multi-modal learning by interleaving cross-attention layers into a pretrained large language model to process visual inputs.

Qwen2-VL-7b [41], based on the Qwen2 architecture, supports image understanding across multiple resolutions and aspect ratios, achieving joint understanding and reasoning through deep fusion of a visual encoder and a language model.

LLaVA-NeXT-7b/8b [42] is an improved version of the LLaVA series that enhances multi-modal reasoning performance through improved visual instruction tuning and higher-resolution image processing.

LLaVA-13B [54] connects a pretrained CLIP [31] visual encoder with the LLaMA language model and employs a two-stage instruction tuning strategy for vision–language alignment.

PaLM-E [21] series (12B/66B/562B) are embodied multi-modal language models that inject visual features directly into the language model input, demonstrating varying levels of reasoning capability on VQA tasks with different parameter scales.

4.4. Implementation Details

The experiments are conducted on an Ubuntu 24.04.1 system using an NVIDIA RTX 4090 GPU, based on Python 3.8 and the PyTorch 1.10.0 deep learning framework. Model training employs the AdamW [55] optimizer for parameter optimization, with the learning rate decaying by a factor of 0.75 per epoch.

The parameter settings in Table 1 are determined as follows. The hidden dimension (768) and the number of attention heads (12) are structurally determined by the RoBERTa-base [34] encoder adopted in our framework, and the dropout rate (0.2) follows its default configuration. The contrastive projection dimension d_p = 256 is consistent with standard practice in CLIP [31] and ALBEF [32]. The pre-alignment layer counts $N_{VS} = N_{QS} = 5$ follow the configuration in [25], and $N_{QV} = 5$ is selected based on the sensitivity analysis in Section 4.7. For training, the learning rate (1×10^{-5}) falls within the recommended range for fine-tuning pretrained Transformers [32,34], the weight decay (1×10^{-4}) follows the AdamW [55] default, and the batch size (32) represents the maximum feasible size on our hardware while being consistent with comparable systems [5,13]. The loss balancing coefficients α, β, and τ are determined through the sensitivity analysis in Section 4.7, while $\lambda = 1.0$ reflects the standard equal-weighting convention for generation and auxiliary losses. The maximum knowledge length (k_max_len = 10) and question length (q_max_len = 15) are set to cover approximately 95% of training samples, the retrieval count (k_max_num = 10) and answer frequency threshold (min_occurrence = 3) follow established KB-VQA conventions [5,13], and the number of training epochs (10) is determined by validation convergence.

Table 1. Experimental Parameter Settings.

Parameter	Value	Description
Epoch	10	Number of training iterations
Dropout	0.2	Dropout Rate of Attention Module
min_occurrence	3	Minimum Answer Frequency Threshold
hidden_size	768	Hidden Dimension of Feed-Forward Network (FFN)
num_heads	12	Number of Heads for Multi-Head Attention (MHA)
N_{VS}	5	Number of Layers of VACE Cross-Modal Encoder
N_{QS}	5	Number of Layers of QACE Cross-Modal Encoder
N_{QV}	5	Number of Layers of CRGF Cross-Modal Encoder (with Cross-Residual Gating)
d_p	256	Output Dimension of VACL Projection Head
k_max_num	10	Maximum Number of Fused Knowledge Entries
k_max_len	10	Maximum Length of Knowledge Triples
q_max_len	15	Maximum Question Length
Batch_size	32	Batch Size
Weight_decay	1×10^{-4}	Weight Decay
lr	1×10^{-5}	Learning Rate (LR)
τ	0.07	Temperature Hyperparameter of VACL
α	0.1	Balancing Coefficient for Contrastive Loss
β	0.01	Balancing Coefficient for Information Bottleneck Loss
λ	1.0	Balancing Coefficient for Semantic Alignment Loss

The detailed parameter settings of the BRIDGE framework are shown in Table 1.

4.5. Comparison with Existing Methods

Table 2 presents the quantitative comparison between BRIDGE and current mainstream methods on the OK-VQA dataset. To comprehensively evaluate the performance of the proposed method, the compared methods cover three technical approaches: traditional retrieval-based methods, LLM prompt-based methods, and end-to-end multi-modal large models. The retrieval-alignment modules of the BRIDGE framework remain unchanged, and only the backbone model of the reader (SAMR) is replaced to verify the generality of the framework.

Table 2. Performance comparison on the OK-VQA dataset.

Category	Model	Explicit Knowledge	Implicit Knowledge	Params	Acc (%)
Retrieval-Based Methods	ConceptBERT [8]	ConceptNet	ViLBERT	~221M	33.66
	KRISP [47]	Wikipedia, ConceptNet	Multi-modal BERT	~200M	38.9
	RVL [9]	Wikipedia, ConceptNet	LXMERT	~240M	39.04
	VRR [46]	ConceptNet, Google Search	—	~200M	39.2
	MAVEx [48]	Wikipedia, ConceptNet, Google Images	—	~221M	40.28
	UnifER+ViLT [3]	ConceptNet	ViLT	~115M	42.13
	KAT-T5 [6]	Wikipedia	T5-large	770M	44.3
	LaKo+T5 [49]	ConceptNet, DBPedia, WebChild	T5	770M	47.01
	TriG [5]	Wikipedia	T5-large	770M	50.5
	RA-VQA [13]	Wikipedia	T5-large	770M	54.5
	RA-VQA-v2 [14]	Wikipedia	T5-XL	3B	62.1
LLM Prompt-Based Methods	KAT-base [6]	Wikipedia	GPT-3	175B	40.93
	DEDR+MM-FiD [4]	Wikipedia	GPT-3, LXMERT	175B	44.57
	PICa-Full [16]	—	GPT-3	175B	48
	KAT (Single) [6]	Wikipedia	GPT-3	175B	53.1
	KAT (Ensemble) [6]	Wikipedia	GPT-3	175B	54.4
	REVIVE [15]	Wikipedia	GPT-3	175B	56.6
	PromptCap [18]	—	GPT-3	175B	60.4
	MM-Reasoner [50]	—	GPT-4, i-Code v2	—	60.4
	Prophet [17]	—	GPT-3	175B	61.2
	ASB [19]	—	LLaMA 2	13B	61.2
	SKP [51]	—	Vicuna-7b	7B	63.3
	PrQAC [52]	—	LLaMA-3	8B	62.89
	GC-KBVQA [53]	—	LLaMA-3	8B	54.57
End-to-End Multi-Modal Large Models	Flamingo [20]	—	Flamingo	80B	57.8
	Qwen2-VL-7b [41]	—	Qwen2-VL-7b	7B	58.8
	LLaVA-NeXT-7b [42]	—	LLaVA-NeXT-7b	7B	58.7
	PaLM-E-12B [21]	—	PaLM-E	12B	60.1
	LLaVA-NeXT-8b [42]	—	LLaVA-NeXT-8b	8B	62.2
	PaLM-E-66B [21]	—	PaLM-E	66B	62.9
	LLaVA-13B [54]	—	LLaVA-13B	13B	64.7
	PaLM-E-562B [21]	—	PaLM-E	562B	66.1
Ours	BRIDGE (BLIP)	Wikipedia, Google Search	BLIP	4B	64.2
	BRIDGE (BLIP)	Wikipedia	BLIP	4B	63.0
	BRIDGE (Qwen2-VL)	Wikipedia, Google Search	Qwen2-VL-7b	7B	66.2
	BRIDGE (LLaVA-NeXT)	Wikipedia	LLaVA-NeXT-8b	8B	66.5
	BRIDGE (LLaVA-NeXT)	Google Search	LLaVA-NeXT-8b	8B	65.4
	BRIDGE (LLaVA-NeXT)	Wikipedia, Google Search	LLaVA-NeXT-8b	8B	**67.8**

Notes: MM-Reasoner [50] is based on GPT-4 and i-Code v2; however, since the number of parameters for these models has not been officially disclosed, they are not listed here. The retrieval-alignment module of the BRIDGE framework remains unchanged; only the backbone model of the reader (SAMR) has been replaced. The total number of parameters is the sum of those in the reader model and the BRIDGE alignment module. Bold values denote the results of our proposed method.

Compared with traditional retrieval-based methods (ConceptBERT [8], KRISP [47], TRiG [5], etc.), the advantage of BRIDGE primarily stems from the multi-source semantic anchors provided by MSAC. These methods typically rely on only a single textual query for knowledge retrieval, whereas BRIDGE constructs a more comprehensive query representation by integrating captions, tags, and OCR text. Even under the Wikipedia-only setting, BRIDGE (BLIP, Wiki only) achieves 63.0%, outperforming RA-VQA [13] (54.5%) and TRiG [5] (50.5%)—both of which also rely solely on Wikipedia—by 8.5 and 12.5 per-

centage points, respectively, confirming that the gains primarily stem from the proposed alignment architecture rather than additional knowledge source coverage. This significant improvement is attributed to the symmetric pre-alignment mechanism of VACE and QACE that effectively bridges the cross-modal semantic gap, as well as the cross-residual gating mechanism of CRGF that substantially suppresses noise introduced during retrieval through semantic residual-driven dynamic gating.

Compared with LLM prompt-based methods (Prophet [17], PromptCap [18], SKP [51], etc.), BRIDGE achieves comparable or even superior performance with substantially fewer parameters. With BLIP as the reader, BRIDGE (64.2%, 4B) surpasses Prophet (61.2%, 175B) and PromptCap [18] (60.4%, 175B) by 3.0 and 3.8 percentage points, respectively, while using less than 2.3% of their parameter count. This improvement is attributed to the cross-residual gating mechanism of CRGF that effectively suppresses fusion noise, and the cross-modal contrastive learning of VACL that enhances alignment quality at the representation space level, compensating for the gap in model scale.

Compared with end-to-end multi-modal large models (LLaVA, Qwen2-VL [41], PaLM-E [21], etc.), BRIDGE consistently outperforms the corresponding base models when using readers of the same scale. BRIDGE (Qwen2-VL [41]) (66.2%) surpasses the original Qwen2-VL-7b [41] (58.8%) by 7.4 percentage points, and BRIDGE (LLaVA-NeXT [42]) (67.8%) surpasses the original LLaVA-NeXT-8b [42] (62.2%) by 5.6 percentage points. This consistent improvement indicates that the retrieval-alignment modules of BRIDGE provide effective knowledge enhancement and semantic alignment support for different reader backbones. Although PaLM-E-562B [21] (66.1%) achieves high accuracy with its massive 562B parameter count, BRIDGE (LLaVA-NeXT [42]) surpasses it with only approximately 8B parameters, less than 1.4% of the former.

Parameter efficiency analysis. The BRIDGE framework demonstrates significant advantages in parameter efficiency. With BLIP as the reader, the total parameter count is approximately 4B, constituting only 2.3% of the GPT-3 (175B) used by Prophet [17] and 0.7% of PaLM-E-562B [21], yet it achieves 64.2% accuracy on OK-VQA. Even with LLaVA-NeXT-8b [42] as the reader (8B), BRIDGE remains far below the parameter counts of PromptCap [18] (175B) and REVIVE [15] (175B) while surpassing them in accuracy by 7.4 and 11.2 percentage points, respectively. This result strongly demonstrates that, without relying on extreme parameter scales, a high-performance visual question answering system can still be constructed in resource-constrained scenarios through refined cross-modal semantic alignment design (MSAC + VACE/QACE pre-alignment + CRGF gated fusion + VACL contrastive learning). Moreover, the consistent performance improvement of BRIDGE when equipped with different readers (BLIP 64.2% $\rightarrow$ Qwen2-VL 66.2% $\rightarrow$ LLaVA-NeXT 67.8%) indicates that the proposed retrieval-alignment modules possess strong generality and can work synergistically with generative models of varying scales.

To further disentangle the contribution of the proposed architecture from the benefit of incorporating dual knowledge sources, we conduct a knowledge source ablation study. As shown in Table 2, when Wikipedia is used as the sole knowledge source, BRIDGE (BLIP, Wiki only) achieves 63.0% on OK-VQA, and BRIDGE (LLaVA-NeXT, Wiki only) achieves 66.5%; introducing Google Search on top of this yields only an additional gain of 1.2 to 1.3 percentage points. This result indicates that the architectural design itself is the primary driver of the overall performance improvement. Taking LLaVA-NeXT as an example, the BRIDGE architecture alone raises accuracy from 62.2% to 66.5%, a gain of 4.3 percentage points, whereas the addition of the extra knowledge source contributes only 1.3 percentage points (66.5% to 67.8%), with the former exceeding the latter by more than a factor of three.

Table 3 reports the comparison between BRIDGE and existing methods on the FVQA benchmark dataset.

Table 3. Performance Comparison on the FVQA Dataset.

Model	Top-1 Acc (%)	Top-3 Acc (%)
BAN [56]	35.69	—
Top1-QQmapping [1]	52.56	59.72
Top3-QQmapping [1]	56.91	64.65
UnifER+LXMERT [3]	51.83	66.83
UnifER+ViLT [3]	55.04	69.72
RVL [9]	54.27	—
DEDR+MM-FiD [4]	61.8	—
BRIDGE (BLIP)	**66.53**	**79.87**

Notes: Bold values denote the results of our proposed method.

BRIDGE achieves the best results on both Top-1 and Top-3 accuracy, reaching 66.53% and 79.87%, respectively. Compared with the previous best method, DEDR+MM-FiD [4], BRIDGE achieves a 4.73 percentage point improvement in Top-1 accuracy while substantially leading UnifER+ViLT [3] by 10.15 percentage points in Top-3 accuracy. The significant leap in the Top-3 metric validates the effectiveness of the MSAC multi-source semantic anchor mechanism in the knowledge retrieval stage: by uniformly transforming heterogeneous visual information into fine-grained semantic symbols, DKR successfully retrieves more relevant candidate knowledge. Meanwhile, the cross-residual gating mechanism of CRGF dynamically filters the inevitable modality noise during retrieval, achieving a breakthrough in Top-1 precision.

Table 4 reports the comparison between BRIDGE and mainstream methods on the A-OKVQA benchmark dataset.

Table 4. Performance comparison on the A-OKVQA dataset.

Model	Direct Answer Acc (%)	Multiple Choice Acc (%)
ClipCap [57]	44	18.1
Pythia [58]	49	25.2
ViLBERT [23]	49.1	30.6
LXMERT [24]	51.4	30.7
KRISP [47]	51.9	33.7
GPV-2 [45]	60.3	48.6
PromptCap+GPT-3 [18]	73.2	56.3
Prophet [17]	76.4	58.2
ASB [19]	—	58.6
SKP [51]	—	65.3
BRIDGE (LLaVA-NeXT)	**76.8**	**66.4**

Notes: Bold values denote the results of our proposed method.

BRIDGE achieves 76.8% under the DA setting, surpassing Prophet [17] (76.4%), and 66.4% under the MC setting, surpassing SKP [51] (65.3%). The consistent performance across datasets validates the generalization capability of the BRIDGE framework under different knowledge reasoning scenarios, demonstrating that the multi-source bridging mechanism of semantic anchors and the representation space alignment established through cross-modal contrastive learning are not overfitting designs tailored to a specific dataset but rather universally applicable cross-modal reasoning enhancement strategies.

Discussion on cross-method comparability. The methods in Tables 2–4 span fundamentally different paradigms (retrieval-based, prompt-based, and end-to-end) and inevitably differ in backbone scale and knowledge source configuration—a characteristic common to KB-VQA benchmarking rather than a limitation of our evaluation. The cross-method tables are therefore intended to situate BRIDGE within the broader performance

landscape. To isolate the contribution of the proposed architecture, the controlled ablation studies in Tables 5–7 fix the backbone (BLIP), knowledge sources, and training protocol, varying only the module under evaluation. We further report a knowledge source ablation in Table 2: under the Wikipedia-only setting, BRIDGE (LLaVA-NeXT) achieves 66.5% on OK-VQA, substantially outperforming TRiG [5] (50.5%) and RA-VQA [13] (54.5%) under the same knowledge source, confirming that the architectural innovations are the principal drivers of the observed improvements.

Table 5. Computational efficiency comparison on OK-VQA. All metrics measure the full inference pipeline from a raw image and question to the final answer.

Method	Params	FLOPs (G) [1]	Latency (ms)	Acc (%)
TRiG [5]	770M	335	106	50.5
RA-VQA [13]	770M	335	106	54.5
SKP [51]	7B	2500	292	63.3
LLaVA-NeXT-8b [42]	8B	2720	313	62.2
BRIDGE (BLIP)	~4B	711	186	64.2
BRIDGE (Qwen2-VL)	~7B	2641	382	66.2
BRIDGE (LLaVA-NeXT)	~8B	2911	419	**67.8**

Notes: [1] FLOPs are computed via fvcore on a single NVIDIA RTX 4090 with batch size 1. Latency is the mean wall-clock time over 100 samples after 10 warm-up iterations, measured using torch.cuda.Event. All methods are evaluated under identical input specifications (image resolution 800 × 600 for VinVL-based pipelines, 336 × 336 for ViT-based pipelines; question length 15; knowledge length 10). KV-cache is enabled for autoregressive decoding, averaged over the OK-VQA mean answer length of 2 tokens. Knowledge base embeddings are pre-indexed with FAISS; the online retrieval cost is included. Baseline implementations: TRiG, RA-VQA, and LLaVA-NeXT are locally deployed from official repositories; SKP is measured using the checkpoint released by the authors. Bold values denote the results of our proposed method.

Table 6. Core component ablation results (BLIP backbone).

Model	OK-VQA Acc (%)	FVQA Top-1 (%)	FVQA Top-3 (%)	A-OKVQA DA (%)
Full Model	64.2	66.53	79.87	70.82
w/o Cross-Residual Gating	63.46	65.82	79.15	70.04
w/o Pre-Alignment (VACE+QACE)	62.87	65.31	78.62	69.18
w/o $\mathcal{L}_{con}$	63.05	65.44	78.71	69.37
w/o $\mathcal{L}_{IB}$	63.82	66.18	79.52	70.48
w/o Semantic Anchors (MSAC)	58.23	60.67	73.28	63.71
Baseline	53.19	56.31	68.42	57.46

Table 7. Semantic anchor component ablation (OK-VQA Dataset, BLIP backbone).

Semantic Anchor Component	OK-VQA Acc (%)	Δ
Caption + Tags + OCR	64.2	—
Caption + Tags	63.27	−0.93
Caption + OCR	62.58	−1.62
Tags + OCR	61.87	−2.33
Only Caption	60.94	−3.26
Only Tags	59.68	−4.52
Only OCR	57.41	−6.79

Table 5 reveals three findings regarding computational efficiency. First, the alignment modules of BRIDGE (VACE, QACE, CRGF, RoBERTa encoding, and dense retrieval) contribute 48 GFLOPs, accounting for 6.8% of BRIDGE (BLIP) and 1.7% of BRIDGE (LLaVA-NeXT). The proposed cross-modal alignment architecture thus imposes minimal overhead

relative to the reader backbone. Second, BRIDGE (BLIP) achieves 64.2% accuracy at 711 GFLOPs—less than one-third of SKP (2500 G, 63.3%) and approximately one-quarter of LLaVA-NeXT-8b (2720 G, 62.2%)—yielding a favorable accuracy-efficiency trade-off for resource-constrained deployment. Third, augmenting LLaVA-NeXT-8b with BRIDGE incurs a 7.0% increase in FLOPs (2720 $\rightarrow$ 2911 G) and a 5.6 percentage point accuracy improvement (62.2 $\rightarrow$ 67.8%), confirming that the retrieval-alignment pipeline provides substantial gains at marginal computational cost. The latency increase (33.9%) exceeds the FLOPs increase because preprocessing modules (VinVL, Oscar, Umi-OCR) execute as sequential pipeline stages; in deployment scenarios with pre-indexed visual features, this sequential overhead is eliminated.

4.6. Ablation Studies

4.6.1. Core Component Ablation

To verify the effectiveness of each core component in the BRIDGE framework, we design a series of ablation experiments on the three benchmark datasets. Specifically, starting from the complete model (Full Model), each key component is sequentially removed to evaluate its contribution. The variants are defined as follows.

- **w/o Cross-Residual Gating**: CRGF is degraded to a standard Q-V cross-modal encoder by removing the cross-residual gated modulation operations (semantic residual computation and gated scaling) after each encoder block layer, retaining only the standard CA $\rightarrow$ SA $\rightarrow$ FFN stacking structure.
- **w/o Pre-Alignment (VACE+QACE)**: The Visual–Anchor Cross-Modal Encoder (VACE) and Question–Anchor Cross-Modal Encoder (QACE) are removed. The linearly projected visual features $W_v F_{vis}$ and question features F_q are directly fed into CRGF for Q-V fusion without the pre-alignment bridging through semantic anchors.
- **w/o $\mathcal{L}_{con}$**: The Visual–Anchor Contrastive Learning module (VACL) is removed, and the training objective does not include the cross-modal contrastive learning loss.
- **w/o $\mathcal{L}_{IB}$**: The Variational Information Bottleneck regularization (VIB) is removed. The answer representation $f_q^{[CLS]}$ is directly used for classification and retrieval without stochastic constraints.
- **w/o Semantic Anchors (MSAC)**: The entire Multi-Source Semantic Anchor Construction module is removed, and no captions, tags, or OCR text are used. In this case, VACE and QACE become simultaneously ineffective due to the absence of anchor inputs; the semantic residuals in CRGF degenerate to zero without the F_{sem} reference, and VACL cannot compute the contrastive loss without z_{sem}. This variant is equivalent to retaining only the visual encoder + question encoder + standard Q-V encoder.
- **Baseline**: Only the visual encoder and question encoder are retained, with answers predicted through simple feature concatenation and an MLP classifier, without using any cross-modal alignment mechanisms, semantic anchors, or auxiliary loss functions.

Table 6 reports the ablation results of each core component on the three benchmark datasets. The removal of each component leads to a consistent performance decline, and the contribution trends remain highly consistent across datasets, validating the generality of the proposed designs.

At the feature representation level, the semantic anchor mechanism (MSAC) contributes the most significant performance gain: removing it causes a 5.97 percentage point accuracy drop on OK-VQA, along with drops of 5.86 (Top-1) and 7.11 (DA) percentage points on FVQA and A-OKVQA, respectively, confirming the core value of multi-source semantic anchors as a cross-modal bridging hub. On this basis, the removal of the pre-alignment stage (VACE+QACE) leads to further performance degradation (OK-VQA -1.33%, FVQA Top-1 -1.22%, A-OKVQA DA -1.64%), indicating that skipping

pre-alignment and directly performing Q-V fusion causes visual and question features to interact ineffectively due to the lack of semantic bridging, thereby validating the necessity of the two-stage design comprising pre-alignment followed by deep fusion.

At the fusion module level, degrading CRGF to a standard Q-V cross-modal encoder without cross-residual gating results in a 0.74 percentage point drop on OK-VQA and a 0.78 percentage point drop on A-OKVQA DA. Although this margin is less pronounced than those of MSAC and pre-alignment, it is consistently observed across all three datasets, validating that the semantic residual-driven gating mechanism provides additional noise suppression capability on top of the standard encoder block.

At the training objective level, the removal of the cross-modal contrastive learning loss $\mathcal{L}_{con}$ causes a 1.15 percentage point drop on OK-VQA and a 1.45 percentage point drop on A-OKVQA DA, validating the effectiveness of constraining cross-modal alignment at the representation space structure level. The gain from $\mathcal{L}_{IB}$ is relatively modest (OK-VQA -0.38%), but it remains stable across all datasets, indicating that the variational information bottleneck effectively compresses redundant information in the fused representation under different scenarios, providing complementary probabilistic assurance for the deterministic noise reduction achieved by cross-residual gating.

4.6.2. Semantic Anchor Component Ablation

Table 7 shows that the complete combination of all three components significantly outperforms any single component or pairwise combination, validating the synergistic complementary value of multi-source information. From the single-component results, captions (60.94%) contribute the most, as they capture the global scene semantics of the image and provide foundational context for most questions. Tags rank second (59.68%), with their value lying in providing fine-grained entity and attribute information that captions may omit. OCR text alone yields the weakest performance (57.41%), since only a subset of images contains recognizable scene text. The complete anchor (64.2%) exceeds the mean accuracy of the three single components (59.34%) by 4.86 percentage points, indicating significant synergistic effects among the three components.

4.6.3. Contrastive Learning Configuration Ablation

Table 8 compares the performance of different contrastive learning configurations, validating the core hypothesis of BRIDGE. V-Sem pairs (64.2%) outperform V-C pairs (63.58%) and Q-V pairs (63.21%) on the OK-VQA dataset, confirming that multi-source semantic anchors exhibit a stronger semantic association with visual content than single captions. The trend that V-C pairs outperform Q-V pairs is consistent with the findings of prior research, further validating that descriptive text (captions/anchors) is more suitable than interrogative text (questions) as the anchoring target for cross-modal contrastive learning. Notably, the combination of V-Sem and Q-V (63.87%) not only fails to surpass V-Sem alone but causes a 0.33 percentage point performance drop. This indicates that simultaneously employing two contrastive learning objectives introduces redundant gradient signals that interfere with the alignment direction of the representation space.

Table 8. Contrastive learning configuration ablation (BLIP backbone).

Model	OK-VQA Acc (%)	FVQA Top-1 (%)	A-OKVQA DA (%)
V-Sem (caption + tags + OCR)	64.2	66.53	70.82
V-C (Only caption)	63.58	65.82	70.15
Q-V	63.21	65.38	69.64
V-Sem and Q-V	63.87	66.16	70.39
Do not use contrastive learning	63.05	65.44	69.37

4.6.4. Robustness Under Noisy Anchor Inputs

Anchor generators—object detectors and OCR engines—are imperfect in practice; we therefore assess whether the cross-residual gating in CRGF attenuates input-side corruption. Noise is injected at inference only, leaving the trained weights untouched.

Two corruption modes are considered at ratios $p \in \{20\%, 40\%, 60\%\}$: in tag noise, each predicted tag is independently replaced with probability p by a category drawn uniformly from the COCO vocabulary; in OCR noise, each character is independently substituted with probability p by a random alphabetic character. The Full Model is compared against the w/o CRG variant, in which CRGF reduces to a standard cross-modal encoder. Results on OK-VQA are reported in Table 9.

Table 9. Robustness of CRGF under noisy anchor inputs (OK-VQA, BLIP backbone). "Gap" denotes the accuracy difference between Full Model and w/o CRG. Noise is injected at inference time without retraining.

Noise Type	Noise Ratio	Full Model Acc (%)	w/o CRG Acc (%)	Gap (Δ)
—	0% (Clean)	64.2	63.46	0.74
Tag	20%	63.83	62.94	0.89
Tag	40%	63.42	62.36	1.06
Tag	60%	62.94	61.71	1.23
OCR	20%	63.99	63.16	0.83
OCR	40%	63.75	62.83	0.92
OCR	60%	63.47	62.46	1.01

Two trends emerge from Table 9. First, the Full Model exhibits consistently slower accuracy decay than the w/o CRG variant; at $p = 60\%$ tag noise, the Full Model loses 1.26 percentage points, whereas the ablated variant loses 1.75 percentage points, a relative degradation 38.9% larger. Second, the gap between the Full Model and the w/o CRG variant grows monotonically with p under both corruption modes—from 0.74 percentage points on clean inputs to 1.23 percentage points at 60% tag noise and 1.01 percentage points at 60% OCR noise. The gating mechanism thus contributes little when anchors are reliable but engages more strongly as their quality deteriorates.

Two mechanisms underlie this behavior. The multi-source anchor is intrinsically redundant: corruption in any single stream is diluted in F_{sem} by the remaining streams, constraining the noise that reaches the gating reference. Complementarily, the sigmoid function in the gating operation saturates at extreme residuals, bounding the modulation and preventing anomalous anchors from disproportionately perturbing the fused output.

4.7. Hyperparameter Sensitivity Analysis

To evaluate the sensitivity of the BRIDGE framework to key hyperparameters, this paper systematically analyzes the effects of the CRGF layer count N_{QV}, contrastive loss weight α, information bottleneck weight β, and temperature coefficient τ on the OK-VQA dataset. The results are shown in Figure 4.

We investigate the sensitivity of four key hyperparameters and report accuracy on OK-VQA in Figure 4, with $N_{VS} = N_{QS} = 5$ fixed following prior work [25]. As shown in Figure 4a, performance peaks at N_{QV} = 5 (64.20%), with shallow networks underperforming due to insufficient cross-modal refinement and deeper ones exhibiting marginal overfitting. For the contrastive loss weight, Figure 4b shows the best result at α = 0.1, where the contrastive loss provides effective alignment without dominating the total gradient. Regarding the information bottleneck weight, Figure 4c indicates that β = 0.01 achieves optimal performance; notably, this is the most sensitive hyperparameter, as increasing β to 0.1 causes a sharp drop to 62.18% due to excessive compression of critical semantic

information. Similarly, Figure 4d reveals that the optimal temperature is $\tau = 0.07$, balancing discrimination sharpness and gradient stability. Overall, BRIDGE is reasonably robust across hyperparameters, with only β requiring careful tuning (recommended below 0.02).

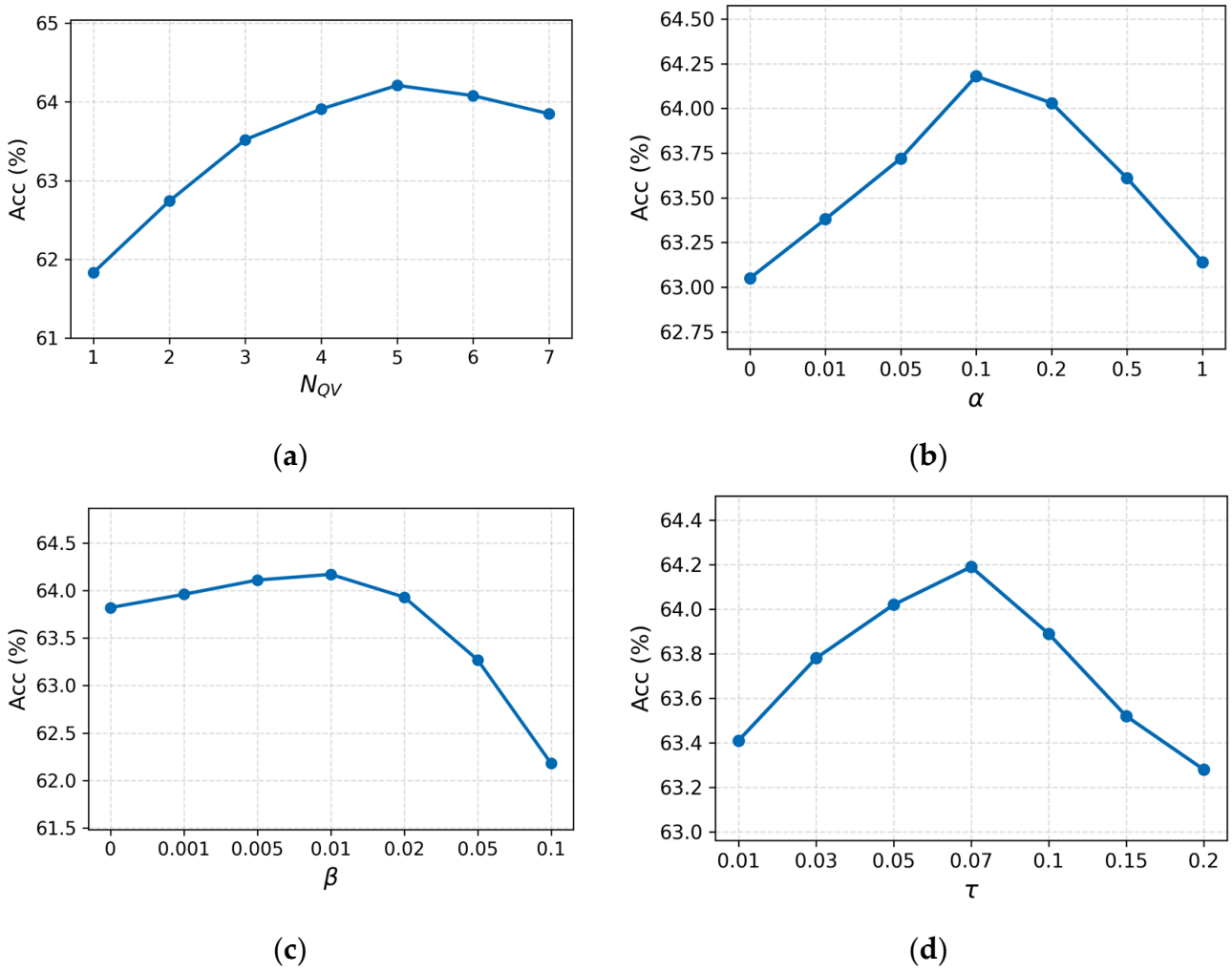


Figure 4. Hyperparameter sensitivity analysis on OK-VQA. (**a**) Impact of the cross-modal refinement depth N_{QV}; (**b**) impact of the contrastive loss weight α; (**c**) impact of the information bottleneck weight β; (**d**) impact of the temperature coefficient τ.

4.8. Qualitative Analysis

To intuitively illustrate the reasoning process of BRIDGE, Figure 5 presents four representative examples, including three correct and one incorrect prediction.

In Figure 5a, the question asks about the contents of visible bottles. The bathroom scene anchor in the caption effectively constrains DKR to retrieve bath-product-related knowledge rather than generic "bottle" knowledge, enabling the model to correctly predict "shampoo." In Figure 5b, the question asks which desk item could help with a cold. Among multiple candidate objects (laptop, book, tissue box), the fine-grained tag "(box, tissue)" provides DKR with a precise retrieval anchor, linking the functional association between "tissue" and "cold" to yield the correct prediction. In Figure 5c, the question asks what a sign on the cart instructs. OCR extracts "Go ahead and push me," and SAMR precisely aligns this action instruction with the intent semantics of "tell you to do" in the question, correctly predicting "push it."

Figure 5d illustrates a typical failure mode: correct knowledge but answer generation drift. DKR accurately retrieves that "orange vehicles are tow trucks used to move aircraft

on the tarmac," yet the reader shifts attention toward the operational object "aircraft" rather than extracting the core functional semantics "transport," producing the incorrect answer "airplane." This reveals that the reader struggles to distinguish functional descriptions from operational objects when both co-occur in the retrieved knowledge.

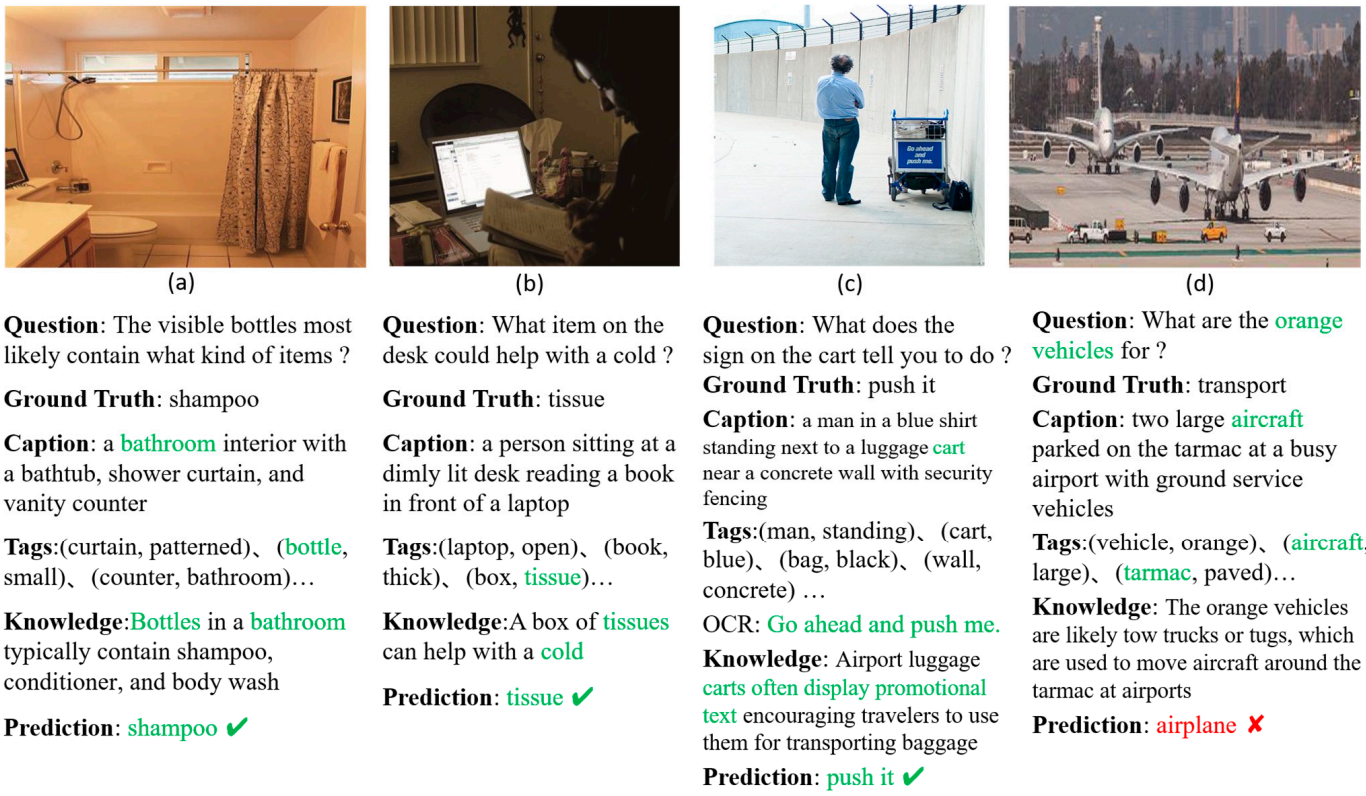


Figure 5. Qualitative analysis of BRIDGE on representative OK-VQA examples. (**a**) Caption-guided retrieval correctly identifies bath products from a bathroom scene; (**b**) fine-grained object tag "(box, tissue)" anchors retrieval to the functionally relevant object; (**c**) OCR-extracted text enables precise alignment with the question intent; (**d**) a failure case where the reader shifts attention from the functional semantics to a co-occurring operational object despite correct knowledge retrieval. Green check marks (✔) indicate correct predictions and red crosses (✘) denote incorrect ones. Colored highlights mark key semantic cues in the retrieved knowledge that contribute to (or mislead) the reasoning process.

Table 10 reveals a clear pattern. In Example (a), where the caption already provides sufficient scene context ("bathroom interior"), all three methods correctly answer "shampoo"—demonstrating that when caption-level information is adequate, even single-caption retrieval (TRiG) and pure visual reasoning (LLaVA-NeXT) can succeed. However, this parity breaks down in Examples (b) and (c), where the critical information lies outside the coverage of the caption. In Example (b), the caption describes the overall scene (person, desk, laptop) but entirely omits the visually inconspicuous tissue box. Without the fine-grained tag "(box, tissue)," the retrieval query of TRiG lacks the key object anchor, leading it to a commonsense guess ("medicine"), while LLaVA-NeXT hallucinates an absent item ("tea"). Only BRIDGE, anchored by the tag, correctly identifies the target. In Example (c), both BRIDGE and LLaVA-NeXT arrive at correct answers but through fundamentally different mechanisms. BRIDGE explicitly extracts the sign text via OCR ("Go ahead and push me") and aligns it with the question through SAMR, producing a complete and faithful answer: "push it." LLaVA-NeXT, benefiting from the visual text recognition capability inherent in its vision encoder, also recognizes part of the text and

outputs "push"—acceptable under the VQA soft voting protocol. However, this success is contingent on the text being clearly visible and in a common font; for degraded, occluded, or multilingual scene text, explicit OCR anchoring provides a more reliable extraction pathway. In contrast, TRiG, which lacks any text recognition capability, resorts to inferring the content of the sign from scene context and produces the functionally plausible but factually incorrect answer "carry luggage."

Table 10. Comparative qualitative analysis of BRIDGE, TRiG [5], and LLaVA-NeXT [42] on representative OK-VQA examples. GT denotes the ground-truth answer. ✔ indicates a correct prediction and ✘ an incorrect one.

Example	Question	GT	BRIDGE	TRiG [5]	LLaVA-NeXT [42]
(a)	The visible bottles most likely contain what kind of items?	shampoo	shampoo ✔	shampoo ✔	shampoo ✔
(b)	What item on the desk could help with a cold?	tissue	tissue ✔	medicine ✘	tea ✘
(c)	What does the sign on the cart tell you to do?	push it	push it ✔	carry luggage ✘	push ✔
(d)	What are the orange vehicles for?	transport	airplane ✘	airplane ✘	airport service ✘

5. Conclusions

In this paper, we propose the BRIDGE framework, which bridges the cross-modal semantic gap in knowledge-enhanced VQA by constructing multi-source semantic anchors (captions, tags, and OCR text) as intermediaries for cross-modal alignment. Specifically, a symmetric pre-alignment mechanism performs bidirectional alignment between semantic anchors and both visual and question features, avoiding direct interaction between heterogeneous modalities. A cross-residual gating fusion module then leverages semantic residuals from complementary modalities to adaptively suppress modality noise, while variational information bottleneck regularization compresses redundant information in the fused representation. Visual–Anchor Contrastive Learning (VACL) module further enhances alignment quality at the representation space level, and dense knowledge retrieval together with a semantic alignment-enhanced generative reader jointly achieve open-domain answer reasoning. Extensive experiments on OK-VQA, FVQA, and A-OKVQA demonstrate that BRIDGE outperforms state-of-the-art methods, and ablation studies confirm the independent contribution of each module, with the multi-source semantic anchors and symmetric pre-alignment providing the most significant gains. One limitation lies in the quality dependency on the underlying anchor generation tools (captioner, detector, OCR); future work plans to incorporate stronger visual foundation models to improve anchor quality. Another limitation is the exclusive reliance on unstructured textual knowledge, which restricts multi-hop reasoning over complex logical chains; we plan to explore combining knowledge graph reasoning with dense retrieval to further enhance the model's deep reasoning capability.

Author Contributions: Conceptualization, F.Z. and Y.H.; Methodology, J.H.; Software, J.H.; Validation, Y.H.; Formal analysis, J.H.; Resources, F.Z.; Data curation, J.H.; Writing—original draft, J.H.; Writing—

review and editing, J.Z., F.Z. and Y.H.; Supervision, J.Z. and Y.H.; Funding acquisition, Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 61862006, and the Natural Science Foundation of Guangxi Province, grant number 2020GXNSFAA159074.

Data Availability Statement: The benchmark datasets used in this study—OK-VQA, FVQA, and A-OKVQA—are publicly available from their respective official repositories. The two external knowledge corpora are also publicly accessible: the Wikipedia corpus (OKVQA_passages, 344K entries) is available as part of the M2KR benchmark at https://huggingface.co/datasets/BByrneLab/multi%5C_task%5C_multi%5C_modal%5C_knowledge%5C_retrieval%5C_benchmark%5C_M2KR (accessed on 25 Month 2026); the Google Search corpus (okvqa_full_corpus, 168,306 entries) is available at the public repository of Luo et al. (https://github.com/luomancs/retriever%5C_reader%5C_for%5C_okvqa) (accessed on 25 Month 2026). The three reader backbone models used in BRIDGE are publicly available on Hugging Face: BLIP (https://huggingface.co/Salesforce/blip2-flan-t5-xl) (accessed on 25 Month 2026), Qwen2-VL (https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct-AWQ) (accessed on 25 Month 2026), and LLaVA-NeXT (https://huggingface.co/llava-hf/llama3-llava-next-8b-hf) (accessed on 25 Month 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. VQA: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2425–2433.
2. Wang, P.; Wu, Q.; Shen, C.; Dick, A.; Van Den Hengel, A. FVQA: Fact-based visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 2413–2427. [CrossRef]
3. Guo, Y.; Nie, L.; Wong, Y.; Liu, Y.; Cheng, Z.; Kankanhalli, M. A unified end-to-end retriever-reader framework for knowledge-based VQA. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 2061–2069.
4. Salemi, A.; Altmayer Pizzorno, J.; Zamani, H. A symmetric dual encoding dense retrieval framework for knowledge-intensive visual question answering. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Taipei, Taiwan, 23–27 July 2023; pp. 110–120.
5. Gao, F.; Ping, Q.; Thattai, G.; Reganti, A.; Wu, Y.N.; Natarajan, P. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 5067–5077.
6. Gui, L.; Wang, B.; Huang, Q.; Hauptmann, A.G.; Bisk, Y.; Gao, J. 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, Seattle, WA, USA, 10–15 July 2022; pp. 956–968.
7. Huang, Z.; Wu, J.; Surana, R.; Jain, R.; Yu, T.; Addanki, R.; Arbour, D.; Kim, S.; McAuley, J. Traceable and Explainable Multimodal Large Language Models: An Information-Theoretic View. In Proceedings of the Second Conference on Language Modeling, Montréal, QC, Canada, 7–10 October 2025.
8. Gardères, F.; Ziaeefard, M.; Abeloos, B.; Lecue, F. Conceptbert: Concept-aware representation for visual question answering. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; pp. 489–498.
9. Shevchenko, V.; Teney, D.; Dick, A.; van den Hengel, A. Reasoning over vision and language: Exploring the benefits of supplemental knowledge. In Proceedings of the Third Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN), Kyiv, Ukraine, 19–23 April 2021; pp. 1–18.
10. Vrandečić, D.; Krötzsch, M. Wikidata: A free collaborative knowledgebase. *Commun. ACM* **2014**, *57*, 78–85. [CrossRef]
11. Wu, Q.; Shen, C.; Wang, P.; Dick, A.; van den Hengel, A. Image captioning and visual question answering based on attributes and external knowledge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1367–1381. [CrossRef] [PubMed]
12. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
13. Lin, W.; Byrne, B. Retrieval augmented visual question answering with outside knowledge. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 11238–11254.

14. Lin, W.; Chen, J.; Mei, J.; Coca, A.; Byrne, B. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 22820–22840.
15. Lin, Y.; Xie, Y.; Chen, D.; Xu, Y.; Zhu, C.; Yuan, L. Revive: Regional visual representation matters in knowledge-based visual question answering. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 10560–10571.
16. Yang, Z.; Gan, Z.; Wang, J.; Hu, X.; Lu, Y.; Liu, Z.; Wang, L. An empirical study of gpt-3 for few-shot knowledge-based vqa. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 3081–3089.
17. Shao, Z.; Yu, Z.; Wang, M.; Yu, J. Prompting large language models with answer heuristics for knowledge-based visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 14974–14983.
18. Hu, Y.; Hua, H.; Yang, Z.; Shi, W.; Smith, N.A.; Luo, J. Promptcap: Prompt-guided image captioning for vqa with gpt-3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 2963–2975.
19. Xenos, A.; Stafylakis, T.; Patras, I.; Tzimiropoulos, G. A simple baseline for knowledge-based visual question answering. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 14871–14877.
20. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: A visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 23716–23736.
21. Driess, D.; Xia, F.; Sajjadi, M.S.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. PaLM-E: An embodied multimodal language model. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 8469–8488.
22. Bender, E.M.; Koller, A. Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 5185–5198.
23. Lu, J.; Batra, D.; Parikh, D.; Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 13–23.
24. Tan, H.; Bansal, M. 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), Hong Kong, China, 3–7 November 2019; pp. 5100–5111.
25. Wang, B.; Ma, Y.; Li, X.; Gao, J.; Hu, Y.; Yin, B. Bridging the cross-modality semantic gap in visual question answering. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *36*, 4519–4531. [CrossRef]
26. Zhang, Y.; Wang, T.; Xue, L.; Lian, W.; Tao, R. ORSI Salient Object Detection via Progressive Interaction and Saliency-Guided Enhancement. *IEEE Geosci. Remote Sens. Lett.* **2025**, *23*, 6002105. [CrossRef]
27. Zhang, Y.; Xiao, Y.; Zhang, Y.; Zhang, T. Video saliency prediction via single feature enhancement and temporal recurrence. *Eng. Appl. Artif. Intell.* **2025**, *160*, 111840. [CrossRef]
28. Zhang, Y.; Wang, T.; Xiao, Y.; Zhang, T.; Zhang, Y.; Tao, R. SJ-PVC: An efficient perceptual video compression scheme based on adaptive QP and rate-distortion optimization. *IEEE Trans. Consum. Electron.* **2025**, *71*, 706–719. [CrossRef]
29. Zhang, Y.; Zhang, Y.; Kang, Y.; Xiao, Y.; Wang, T. Transformer-Based Multi-Level Semantic Guidance and Multi-Branch Feature Fusion Network for Driver Attention Prediction. *IEEE Sens. J.* **2026**. [CrossRef]
30. Zhang, Y.; Zhen, J.; Sun, S.; Liu, T.; Huo, L.; Wang, T. SCAFNet: A Semantic Compensated Adaptive Fusion Network for Remote Sensing Images Change Detection. *IEEE Geosci. Remote Sens. Lett.* **2026**, *23*, 6003405. [CrossRef]
31. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Hangzhou, China, 27–28 October 2021; pp. 8748–8763.
32. Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; Hoi, S.C.H. Align before fuse: Vision and language representation learning with momentum distillation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 9694–9705.
33. Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; Gao, J. Vinvl: Revisiting visual representations in vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 5579–5588.
34. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
35. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2020; pp. 121–137.
36. HIROI-SORA Umi-OCR Text Recognition Tool, Version 2.1.5. Available online: https://github.com/hiroi-sora/Umi-OCR (accessed on 1 June 2025).

37. Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W.-T. Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 6769–6781.
38. Kim, W.; Son, B.; Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021; pp. 5583–5594.
39. Li, L.H.; Yatskar, M.; Yin, D.; Hsieh, C.J.; Chang, K.W. Visualbert: A simple and performant baseline for vision and language. *arXiv* **2019**, arXiv:1908.03557. [CrossRef]
40. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 12888–12900.
41. Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv* **2024**, arXiv:2409.12191.
42. Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; Li, C. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv* **2024**, arXiv:2407.07895.
43. Oord, A.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
44. Marino, K.; Rastegari, M.; Farhadi, A.; Mottaghi, R. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Proceedings of the IEEE/cvf Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3195–3204.
45. Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; Mottaghi, R. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2022; pp. 146–162.
46. Luo, M.; Zeng, Y.; Banerjee, P.; Baral, C. Weakly-supervised visual-retriever-reader for knowledge-based question answering. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 6417–6431.
47. Marino, K.; Chen, X.; Parikh, D.; Gupta, A.; Rohrbach, M. 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, Nashville, TN, USA, 19–25 June 2021; pp. 14111–14121.
48. Wu, J.; Lu, J.; Sabharwal, A.; Mottaghi, R. Multi-modal answer validation for knowledge-based vqa. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 2712–2721.
49. Chen, Z.; Huang, Y.; Chen, J.; Geng, Y.; Fang, Y.; Pan, J.Z.; Zhang, N.; Zhang, W. Lako: Knowledge-driven visual question answering via late knowledge-to-text injection. In Proceedings of the 11th International Joint Conference on Knowledge Graphs, Hangzhou, China, 27–28 October 2022; pp. 20–29.
50. Khademi, M.; Yang, Z.; Frujeri, F.; Zhu, C. Mm-reasoner: A multi-modal knowledge-aware framework for knowledge-based visual question answering. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 6–10 December 2023; pp. 6571–6581.
51. Wang, Q.; Ji, R.; Peng, T.; Wu, W.; Li, Z.; Liu, J. Soft knowledge prompt: Help external knowledge become a better teacher to instruct llm in knowledge-based vqa. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 11–16 August 2024; pp. 6132–6143.
52. Jiang, P.; Ibrayim, M.; Wang, L.; Xu, W. PrQAC: Prompting LLaMA3 with question-aware image captions and answer candidates for knowledge-based VQA. *Inf. Process. Manag.* **2026**, *63*, 104606. [CrossRef]
53. Moradi, M.M.; Mudur, S. Crafting Descriptive Information for a Zero-shot Method to Improve Knowledge-Based Visual Question Answering Performance. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Tucson, AZ, USA, 27 February–4 March 2026; pp. 3120–3128.
54. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 34892–34916.
55. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
56. Kim, J.H.; Jun, J.; Zhang, B.T. Bilinear attention networks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 1571–1581.
57. Mokady, R.; Hertz, A.; Bermano, A.H. Clipcap: Clip prefix for image captioning. *arXiv* **2021**, arXiv:2111.09734. [CrossRef]
58. Jiang, Y.; Natarajan, V.; Chen, X.; Rohrbach, M.; Batra, D.; Parikh, D. Pythia v0. 1: The winning entry to the vqa challenge 2018. *arXiv* **2018**, arXiv:1807.09956.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.