Context-Aware Knowledge Harmonization for Visual Question Reasoning

Lina Vermeersch *, Quentin Moor, Elodie Fairchild, Sarah Van Steen

Posted Date: 24 November 2025

doi: 10.20944/preprints202511.1720.v1

Keywords: visual question answering; knowledge integration; semantic misalignment; uncertainty modeling; adaptive knowledge control; knowledge graph reasoning

 doi:10.20944/preprints202511.1720.v1

Article

Context-Aware Knowledge Harmonization for Visual Question Reasoning

Lina Vermeersch *, Quentin Moor, Elodie Fairchild and Sarah Van Steen

Université libre de Bruxelles
* Correspondence: lvermeersc@ulb.be

Abstract

Knowledge-intensive visual question answering requires a model to fluidly integrate visual perception, linguistic comprehension, and external knowledge sources. Although recent advances in knowledge-based VQA have explored the incorporation of structured and unstructured knowledge, they frequently overlook the discrepancies between the visual scene and the retrieved knowledge, resulting in semantic conflicts that degrade reasoning quality. To address this long-standing issue, we introduce SENSEALIGN, a unified context-harmonization framework designed to assess, quantify, and mitigate semantic divergence between image-grounded evidence and externally retrieved knowledge. The core principle behind SENSEALIGN is to provide an adaptive mechanism that evaluates whether the newly incorporated knowledge is consistent with the visual–linguistic context, and proportionally adjusts its influence based on a principled inconsistency score. Specifically, we first formulate a novel semantic discrepancy estimator that combines caption-based uncertainty signals with a cross-context semantic similarity evaluation, allowing the system to diagnose whether external knowledge aligns with the underlying visual semantics. Building upon this inconsistency estimator, we further develop an adaptive knowledge assimilation strategy that dynamically regulates explicit knowledge from structured sources and implicit knowledge encoded in pretrained multimodal models. Through this perspective, SENSEALIGN offers a general mechanism for preventing over-reliance on irrelevant or misleading facts while still enabling the model to leverage genuinely helpful knowledge. Comprehensive experiments on the OK-VQA benchmark demonstrate that our approach consistently surpasses strong baselines and establishes a new state-of-the-art performance. These results highlight the significance of explicitly modeling semantic compatibility when integrating heterogeneous knowledge for visual reasoning tasks..

Keywords: visual question answering; knowledge integration; semantic misalignment; uncertainty modeling; adaptive knowledge control; knowledge graph reasoning

1. Introduction

Knowledge-based visual question answering (KVQA) focuses on answering questions whose solutions demand not only an understanding of the visual scene and natural language queries but also the incorporation of complementary external knowledge. Unlike conventional VQA tasks that rely solely on visual recognition and textual inference, KVQA aspires to approximate human-level reasoning by leveraging background knowledge about concepts, attributes, events, or commonsense relations that extend beyond the pixels of the image. This enriched reasoning paradigm, while powerful, introduces new challenges stemming from the heterogeneity and noisiness of large-scale knowledge sources. As highlighted in prior works exploring multi-modal integration frameworks [1,2], the process of injecting knowledge through entity-level retrieval or keyword grounding often retrieves overly generic or contextually inappropriate information, thereby introducing contradictions or semantic drift relative to the true image content.

 doi:10.20944/preprints202511.1720.v1

To illustrate this persistent challenge, when an image depicts a ripe banana that visually appears green, a knowledge base may nonetheless retrieve widely known facts such as *(banana, HasProperty, yellow)*, resulting in an answer that contradicts the observable visual cues. This discrepancy exemplifies what we term as *semantic inconsistency*, a mismatch between the actual visual–linguistic context and the external knowledge triggered by objects or keywords. Such inconsistencies become especially problematic because traditional retrieval-based KVQA systems tend to assume that all retrieved knowledge is relevant, thereby amplifying the negative impact of incorrect or overly generic information on final answer prediction. Although recent works attempt to refine the process of entity linkage or graph-based knowledge extraction [3–5], they typically overlook the necessity of verifying whether the retrieved knowledge faithfully aligns with the grounded visual context.

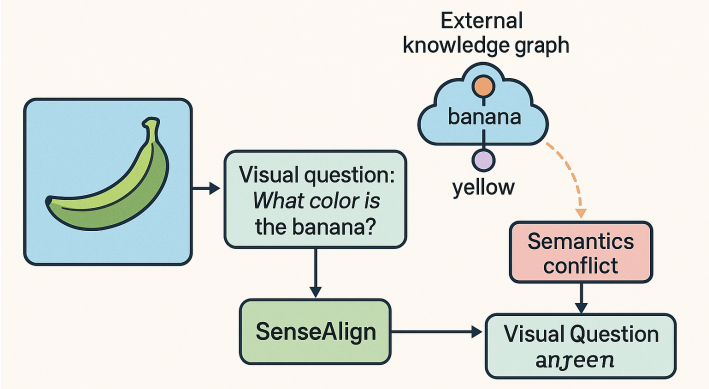

Figure 1. Motivating illustration of semantic inconsistency in knowledge-based visual question reasoning. Although external knowledge suggests that a banana is typically yellow, the visual evidence indicates a green banana. SENSEALIGN detects this mismatch and harmonizes external knowledge with the grounded visual context to produce the correct answer.

Given these observations, it becomes critical to establish a more principled mechanism that explicitly evaluates the compatibility between external knowledge and the visual evidence. Motivated by this need, our work proposes to systematically quantify semantic inconsistency and leverage it to guide knowledge integration. We begin by formulating a measure that exploits caption-generation dynamics, where unusual or ill-fitting knowledge tends to produce captions with higher uncertainty or reduced semantic coherence. By comparing captions generated with different knowledge conditions against the ground-truth description of the image, we obtain a fine-grained signal that reflects whether the introduced knowledge harmonizes with or disrupts the contextual understanding. This enables a robust and interpretable method for identifying misaligned knowledge that may hinder the reasoning process.

Beyond caption uncertainty, we further extend the inconsistency estimation with a contextual similarity assessment that leverages a knowledge context model pretrained on large-scale commonsense corpora. When the knowledge-driven contextual embedding diverges from the visually grounded semantics, the resulting uncertainty and representational gap jointly reflect the level of semantic mismatch. This dual-perspective evaluation—spanning both generative uncertainty and contextual semantic similarity—provides a comprehensive lens for diagnosing misalignment between visual content and retrieved knowledge. Such an estimation is especially advantageous for KVQA because it enables the system to down-weight or suppress misleading knowledge while retaining high-value information that genuinely supports answer prediction.

Building upon this semantic inconsistency model, we further introduce an adaptive knowledge assimilation strategy designed to regulate the degree of external knowledge influence. In traditional KVQA settings, explicit knowledge from structured sources such as relational knowledge graphs and implicit knowledge encoded within multimodal transformers are both useful yet prone to conflict. Our strategy dynamically modulates these sources by increasing reliance on the most contextually consistent components, thereby preventing the accumulation of spurious reasoning patterns. Through

 doi:10.20944/preprints202511.1720.v1

this lens, our approach facilitates smoother integration of heterogeneous knowledge, enabling better alignment between visual perceptions and retrieved facts.

This study contributes to the literature in several significant ways. First, we propose a novel semantic inconsistency estimator rooted in caption-generation uncertainty and semantic similarity comparison, offering a flexible and interpretable measure for diagnosing misaligned knowledge. Second, we introduce SENSEALIGN, an adaptive external knowledge assimilation framework that uses this inconsistency signal to modulate how much explicit and implicit knowledge is injected into KVQA models. Third, by integrating relational graph-based representations and multimodal contextual encoders into the SENSEALIGN pipeline, we establish a more reliable and semantically coherent reasoning process that mitigates the adverse effects of irrelevant knowledge. Finally, extensive experiments on the OK-VQA dataset validate the effectiveness of our approach and demonstrate its superiority over prior state-of-the-art methods, underscoring the value of aligning external knowledge with grounded visual semantics.

In summary, this work advocates for a shift in KVQA research: from indiscriminately incorporating external knowledge to thoughtfully balancing knowledge utility with contextual alignment. Through SENSEALIGN, we show that addressing semantic inconsistency is not merely an auxiliary feature but a central requirement for developing robust and accurate visual question reasoning systems.

2. Related Work

2.1. Pre-Trained Multimodal Foundations for KVQA

A considerable body of research in multimodal representation learning has investigated how visual objects and linguistic units can be jointly encoded to facilitate downstream reasoning tasks. Earlier explorations demonstrated that object-level features extracted from region proposal networks, when tokenized and aligned with textual tokens, can be effectively processed via Transformer-based self-attention mechanisms [6,7]. These models learn cross-modal correspondences that outperform conventional fusion-based architectures [8], especially in tasks requiring fine-grained grounding between image entities and question semantics. Building on this insight, our work also leverages the representational richness of pre-trained multimodal encoders.

Beyond these models, many KVQA pipelines further incorporate features derived from Faster R-CNN, ResNet, or other vision encoders, in combination with question embeddings produced by pretrained language models [9,10]. [9] proposed a Bilinear Attention Map formulation to generate a joint embedding space that reflects subtle visual–textual dependencies. Meanwhile, [10] introduced a 3-way Tucker fusion design enabling complex multiplicative interactions between image regions and question tokens.

Several studies also highlight that pre-trained models can indirectly encode commonsense knowledge, though such implicit knowledge may be incomplete or incongruent with external facts. [1] utilized ArticleNet, which retrieves supplementary information via Wikipedia search APIs triggered by entity-level keywords. Similarly, [2] extracted additional knowledge based on object labels from detection models. While these approaches demonstrate the potential of leveraging implicit and explicit knowledge simultaneously, they often lack mechanisms to assess whether the retrieved knowledge faithfully matches the visual context. The present work advances this by incorporating semantic inconsistency estimation, enabling SENSEALIGN to better regulate the amount and relevance of external knowledge utilized.

2.2. Graph-Structured Knowledge for Visual Reasoning

Graph-based reasoning models have emerged as a parallel line of investigation, motivated by the intuition that structured representations can capture rich entity–relation patterns across modalities [3, 11–14]. [11] proposed the Neural State Machine, adopting a probabilistic graph environment for multi-step visual reasoning. In the context of video-based tasks, [13] constructed a video scene graph augmented with caption generation modules to enhance temporal relational reasoning. [12]

 doi:10.20944/preprints202511.1720.v1

further explored heterogeneous graph alignment networks to incorporate inter- and intra-modality dependencies for video-QA, demonstrating the versatility of graph structures for long-range reasoning.

Other approaches combine heterogeneous features—visual, linguistic, and numeric—into a unified graph and employ node-level aggregation mechanisms [14]. However, these methods typically emphasize the visual content and may struggle when the question requires significant external knowledge. In response, [3] introduced a model that fuses a concept graph derived from external knowledge with an image scene graph, enabling relational alignment between detected objects and knowledge entities. Nonetheless, the scene graph relation types in such models remain limited, and for datasets such as OK-VQA, location-based scene graph construction has shown little improvement in factual reasoning scenarios that require commonsense knowledge.

2.3. Hybrid Pre-Trained and Graph-Enhanced KVQA Pipelines

Several recent studies combine the strengths of pre-trained multimodal encoders and graph reasoning modules in an attempt to capture both implicit conceptual associations and explicit relational structures [4,5,15]. [15] proposed multimodal graph networks aimed at compositional generalization, but their evaluation focused predominantly on object recognition attributes such as shape and count, leaving knowledge-intensive settings underexplored. [4] developed a Knowledge Graph Augmented model that integrates visual features with relational subgraphs constructed from retrieved knowledge. However, their method constructs subgraphs solely from image object labels and question keywords, overlooking whether these elements are semantically compatible with the actual scene.

Similarly, [5] fused representations from a BERT-based encoder with graph-derived concept embeddings. Yet, when inconsistencies arise between the graph-based signals and the multimodal encoder's contextual representation, the injection of external knowledge may mislead the reasoning process rather than improve it. These limitations motivate our introduction of SENSEALIGN, which explicitly quantifies semantic inconsistency at the knowledge–context interface and uses this signal to regulate knowledge assimilation.

2.4. Knowledge Filtering and Conflict Mitigation in VQA

Another emerging research direction concerns the filtering, validation, or reweighting of external knowledge before its incorporation into visual reasoning. Some methods employ heuristic-based filtering strategies, such as ranking retrieved knowledge snippets by lexical relevance or TF-IDF score, but these approaches overlook deeper semantic misalignments that arise when knowledge contradicts what is visually observed. Other techniques explore confidence-driven schemes, treating the retrieval model's confidence as a proxy for usefulness; however, confidence does not reliably capture whether the knowledge is factually incompatible with the current visual question context.

Recent advancements in uncertainty modeling offer a more principled alternative. For instance, approaches leveraging stochastic attention, dropout-based variance estimation, or ensemble-based predictive dispersion have shown potential for detecting anomalous predictions in multimodal inference. Yet, these techniques have rarely been applied to knowledge integration. Our work extends this line of research by leveraging caption-generation uncertainty as a diagnostic tool for semantic conflict, capturing cases where the knowledge-conditioned caption diverges from visually grounded semantics. This provides a finer granularity of conflict detection than traditional retrieval-ranking strategies.

2.5. Semantic Alignment and Contextual Consistency in Knowledge-Intensive AI

Finally, a broader set of studies investigates semantic alignment across modalities and data sources, particularly in tasks requiring cross-domain consistency. Research in commonsense reasoning highlights the importance of ensuring that inferred facts do not contradict observed evidence, especially in multimodal environments with incomplete or ambiguous signals. Knowledge representation studies have proposed embedding-based alignment techniques where the distance between contextual embeddings indicates the degree of compatibility. Similarly, multimodal Transformers incorporate cross-attention mechanisms to reconcile conflicting signals from different modalities. However,

 doi:10.20944/preprints202511.1720.v1

few works explicitly quantify semantic conflict at the intersection of visual grounding and external knowledge.

Motivated by these gaps, our SENSEALIGN framework leverages both semantic similarity measures and uncertainty-driven indicators to construct a unified inconsistency estimator. This estimator serves as a robust foundation for dynamically controlling the injection of external knowledge, ensuring that only contextually coherent and visually compatible facts contribute to downstream reasoning. Through this design, SENSEALIGN advances the alignment problem beyond retrieval relevance and toward a more cognitively grounded notion of semantic harmony between what is seen and what is known.

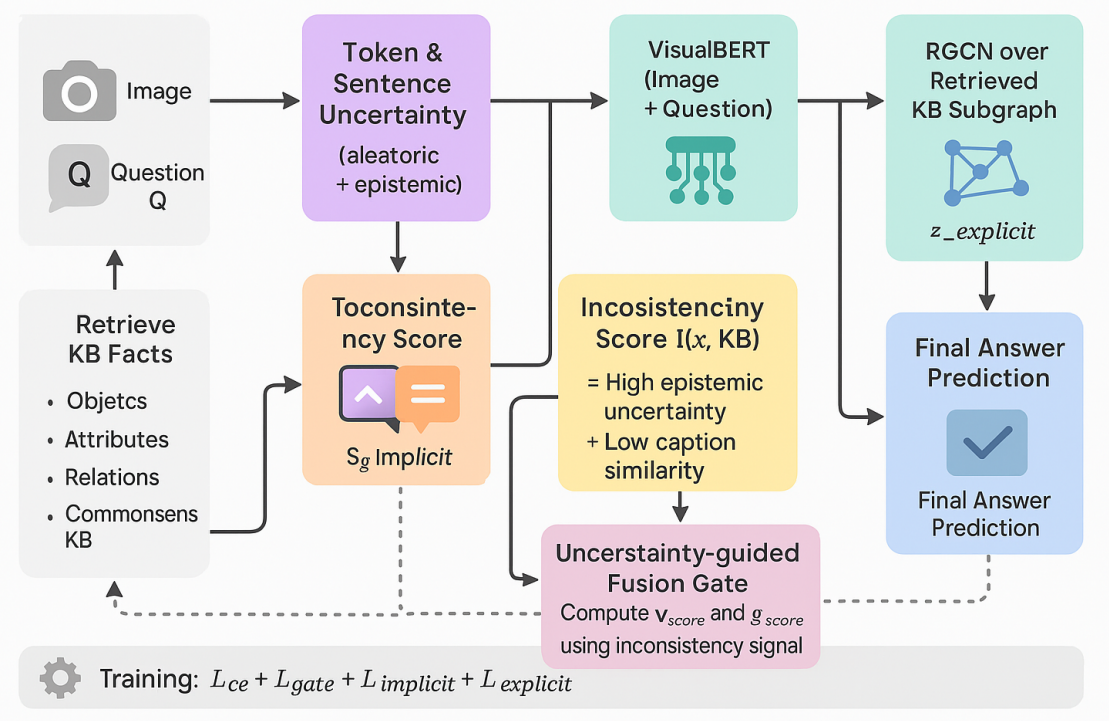

Figure 2. A high-level pipeline of the SenseAlign framework, showing how semantic inconsistency is estimated from KB-conditioned captions and used to adaptively fuse implicit (VisualBERT) and explicit (RGCN) knowledge for KVQA.

3. SenseAlign Framework

In this section, we introduce the proposed SENSEALIGN framework, which explicitly models semantic inconsistency between an image and an external knowledge base (KB) and then uses this signal to regulate the integration of implicit and explicit knowledge for KVQA. We first present how semantic inconsistency is quantified by combining uncertainty modeling and semantic similarity between captions. We then describe how this inconsistency-aware signal is used to adaptively control the contributions of visual–linguistic features and KB-derived features in the final answer prediction process.

3.1. Semantic Inconsistency Between Image Context and External Knowledge

The central idea behind SENSEALIGN is that external knowledge should only be trusted when it is semantically compatible with the visual–linguistic context of a given VQA instance. If the knowledge retrieved from a KB contradicts or weakly correlates with what is grounded in the image and the question, injecting such knowledge can easily lead to hallucinated or biased predictions. To operationalize this intuition, we quantify the degree of mismatch between the image context and the KB by exploiting caption generation. The captioning model is first exposed to the image and the externally retrieved knowledge, and then its predictive behavior is analyzed in terms of uncertainty and similarity to reference captions. Intuitively, if the KB is consistent with the image, the resulting

 doi:10.20944/preprints202511.1720.v1

caption should be confident and semantically close to the reference description of the scene; otherwise, the caption becomes unstable or diverges from the ground-truth semantics.

Inspired by [16], we adopt an uncertainty-based formulation for captioning and extend it to define a semantic inconsistency measure tailored for KVQA. The measure is computed at both token and sentence levels, and then aggregated into a single scalar value that reflects the overall reliability of the KB with respect to the given image and question. This value will later drive the gating mechanism that balances implicit (visual–linguistic) and explicit (KB) knowledge in SENSEALIGN.

3.1.1. Ensemble-Based Uncertainty Estimation for Caption Generation in KVQA

In conventional image captioning, given an input x (e.g., an image or a combination of image and auxiliary information), the goal is to generate a sentence $y = (y_1, \ldots, y_k)$ by learning the conditional distribution $p(y|x)$. Most modern captioning systems model this distribution autoregressively, predicting each token based on the context formed by the input and previously generated tokens. Formally, the conditional distribution factorizes as

$$p(y|x) = p(y_1|x) \prod_{i=2}^{k} p(y_i|x, y_1, \cdots, y_{i-1}), \tag{1}$$

where y_i denotes the i-th token in the sentence and $x, y_1, \ldots, y_{i-1}$ defines the context c_i used to predict the next token.

For a given context c_i, the captioning model defines a probability distribution over a vocabulary V, i.e., $p(y_i = v|c_i)$ for $v \in V$. Semantically plausible tokens (e.g., 'green" for a green banana) tend to concentrate probability mass, whereas implausible words (e.g., 'beach" for an indoor office scene) should ideally receive negligible probability. When the model assigns non-trivial probability to such implausible tokens, we refer to them as hallucinated words. Let $V_h^{(c_i)} \subseteq V$ denote the set of hallucinated words under the context c_i; then the probability mass assigned to hallucinations can be written as

$$p(y_i \in V_h^{(c_i)}) = \sum_{v \in V_h^{(c_i)}} p(y_i = v|c_i). \tag{2}$$

A high value of $p(y_i \in V_h^{(c_i)})$ indicates that the model is tempted to generate contextually irrelevant tokens, which is symptomatic of semantic conflict or model uncertainty.

Uncertainty in token prediction is often quantified using entropy. For each context c_i, the predictive entropy of the token distribution is defined as

$$\begin{aligned} H(y_i|c_i) &= -\sum_{v \in V} p(y_i = v|c_i) \log p(y_i = v|c_i) \\ &= -\sum_{v \in V \setminus V_h^{(c_i)}} p(y_i = v|c_i) \log p(y_i = v|c_i) \\ &\quad -\sum_{v \in V_h^{(c_i)}} p(y_i = v|c_i) \log p(y_i = v|c_i). \end{aligned} \tag{3}$$

The above decomposition reveals two qualitatively different components: (i) uncertainty related to choosing among contextually appropriate tokens; and (ii) uncertainty arising from assigning non-negligible probability to hallucinated tokens. The second component is of particular interest for KVQA, as it is closely related to how incompatible external knowledge can perturb the captioning model.

In a Bayesian view, the overall predictive uncertainty can be further decomposed into aleatoric and epistemic parts [17–19]. Aleatoric uncertainty captures the inherent noise in the data (e.g., occlusions or ambiguous scenes), while epistemic uncertainty stems from model parameters and limited training coverage. To approximate these components, we adopt an ensemble-based modeling strategy [20]. Let $q(w)$ be the posterior distribution over model parameters w, and let $H(y_i|c_i, w)$ be the entropy of the

 doi:10.20944/preprints202511.1720.v1

predictive distribution when the parameters are fixed to w. Aleatoric uncertainty is then approximated by

$$u_{al}(y_i|c_i) = \mathbb{E} * q(w)[H(y_i|c_i, w)] = \frac{1}{M} \sum * m = 1^M H_m(y_i|c_i), \tag{4}$$

where $H_m(y_i|c_i)$ is the entropy computed from the m-th ensemble member and M is the ensemble size. Epistemic uncertainty is estimated by subtracting the aleatoric component from the total predictive entropy:

$$u_{ep}(y_i|c_i) = H(y_i|c_i) - \mathbb{E} * q(w)[H(y_i|c_i, w)] = H(y_i|c_i) - u * al(y_i|c_i). \tag{5}$$

For each caption sequence, we aggregate token-level uncertainties to obtain sentence-level measures. A simple yet effective aggregation is the average across positions:

$$\bar{u} * al(y|x) = \frac{1}{k} \sum * i = 1^k u_{al}(y_i|c_i), \quad \bar{u} * ep(y|x) = \frac{1}{k} \sum * i = 1^k u_{ep}(y_i|c_i). \tag{6}$$

Larger values of $\bar{u}_{ep}(y|x)$ typically indicate that the model is uncertain due to a mismatch between the training distribution and the current input. In our setting, when the caption generator is conditioned on external knowledge that conflicts with the image, epistemic uncertainty tends to increase, thus signaling semantic inconsistency between the KB and the visual context.

A recent line of work suggests that captioning models pre-trained on large-scale datasets implicitly encode a substantial amount of commonsense and factual knowledge [21]. We adopt such a pre-trained captioning backbone as a knowledge-aware generator. By feeding the KVQA images and the associated KB-derived context into this model and then computing $\bar{u} * al(y|x)$ and $\bar{u} * ep(y|x)$ via an ensemble, we obtain a principled uncertainty profile describing how confidently the model can integrate the external knowledge with the visual evidence.

3.1.2. Measuring Semantic Similarity Between Captions

Uncertainty alone does not fully characterize semantic inconsistency, since a model can be confident yet wrong, or uncertain yet still produce a semantically acceptable caption. To complement the uncertainty view, SENSEALIGN also considers the semantic similarity between the knowledge-conditioned caption and a reference description of the image. If the generated caption diverges strongly from the ground-truth caption(s), the utilized knowledge is likely to be misaligned with the true scene.

Let S_g denote a caption generated by the pre-trained model under a particular knowledge configuration (e.g., conditioned on the KB facts associated with the detected objects and question keywords), and let S_t denote a ground-truth caption describing the same image. We employ the Sentence-BERT (S-BERT) encoder [22] to map each sentence into a dense representation $f(S) \in \mathbb{R}^d$. The semantic similarity between the two captions is then computed using cosine similarity:

$$sim^{cap}(S_g, S_t) = \frac{f(S_g) \cdot f(S_t)}{|f(S_g)| \cdot |f(S_t)|}, \quad f : \text{encoder}. \tag{7}$$

The value $sim^{cap}(S_g, S_t)$ lies in $[-1, 1]$, where higher values indicate greater semantic alignment.

In practical KVQA datasets, each image may have multiple reference captions $S_t^{(1)}, \ldots, S_t^{(L)}$. To robustly evaluate similarity, we aggregate over all reference captions, for example by taking the maximum or average similarity:

$$\begin{aligned} sim^{cap} * \max(S_g) &= \max * \ell \in 1, \ldots, L sim^{cap}(S_g, S_t^{(\ell)}), \\ sim^{cap} * \text{avg}(S_g) &= \frac{1}{L} \sum * \ell = 1^L sim^{cap}(S_g, S_t^{(\ell)}). \end{aligned} \tag{8}$$

In our implementation, we primarily use $sim^{cap}_{\max}(S_g)$, as it allows the generated caption to match any one of the plausible references and thus better accommodates linguistic variability.

 doi:10.20944/preprints202511.1720.v1

When external knowledge introduces incorrect assumptions (e.g., forcing the caption to mention "yellow" for a visually green banana), the resulting $sim^{cap}_{\max}(S_g)$ tends to decrease. By monitoring this similarity in conjunction with uncertainty, SENSEALIGN can detect cases where the KB is misleading and should be down-weighted in the final reasoning pipeline.

3.1.3. Combining Uncertainty and Similarity into a Semantic Inconsistency Score

To quantify the overall inconsistency between the external KB and the image–question context, we combine the uncertainty and similarity indicators into a single scalar measure. Intuitively, semantic inconsistency should increase when epistemic uncertainty is high and caption similarity is low. Let $\bar{u} * ep(y|x)$ denote the sentence-level epistemic uncertainty defined in Eq. (6) and let s^{cap} denote a similarity score (e.g., $sim^{cap} * \max(S_g)$). We define an inconsistency score $I(x, \mathrm{KB})$ as

$$I(x, \mathrm{KB}) = \alpha, \bar{u}_{ep}(y|x) + \beta, (1 - s^{cap}), \tag{9}$$

where $\alpha, \beta \geq 0$ are hyperparameters that balance the contributions of uncertainty and similarity. In practice, these coefficients can be tuned on a validation set or normalized so that each term occupies a comparable numerical range.

For downstream usage, we also normalize the inconsistency score into a bounded range, e.g.,

$$\tilde{I}(x, \mathrm{KB}) = \frac{I(x, \mathrm{KB}) - \mu_I}{\sigma_I}, \tag{10}$$

where μ_I and σ_I are the mean and standard deviation of $I(x, \mathrm{KB})$ over the training set. The normalized score $\tilde{I}$ can be further transformed into a consistency confidence by applying a sigmoid operation:

$$\gamma(x, \mathrm{KB}) = \sigma(-\tilde{I}(x, \mathrm{KB})), \tag{11}$$

where larger γ implies higher trust in the KB. This scalar γ becomes a key control signal in the SENSEALIGN framework, guiding how much attention should be allocated to explicit knowledge versus implicit visual–linguistic cues.

3.2. Uncertainty-Aware Knowledge-Based Visual Question Answering

Building upon the semantic inconsistency modeling described above, we now present how SENSEALIGN fuses implicit and explicit knowledge for KVQA. The core idea is to treat the inconsistency score as a dynamic gate that modulates the relative importance of KB-derived features and multimodal contextual representations. When semantic inconsistency is high (large $\tilde{I}$, small γ), the model relies more on image–question information; when inconsistency is low, the KB is trusted more strongly.

Formally, SENSEALIGN relies on two complementary knowledge sources [5]: (i) explicit knowledge, extracted from an external KB using a relational graph convolutional network (RGCN); and (ii) implicit knowledge, encoded in a visual–linguistic Transformer such as VisualBERT. The following subsections describe how uncertainty and similarity are used to regulate these two components.

3.2.1. Uncertainty-Guided Knowledge Utilization

To operationalize the gating mechanism, we construct a feature vector from the semantic similarity and uncertainty indicators and feed it into a small neural controller that outputs two scalar scores. Let $sim^{cap} \in \mathbb{R}$ denote the similarity measure (e.g., $sim^{cap} * \max$), and let $u \in \mathbb{R}$ be a suitable aggregation of uncertainty (for instance, a weighted sum of $\bar{u} * al$ and $\bar{u}_{ep}$). We concatenate these two values and

 doi:10.20944/preprints202511.1720.v1

apply a linear projection followed by a sigmoid function to obtain scores for implicit and explicit knowledge streams:

$$\begin{aligned} v^{score} &= \sigma(W_v * [sim^{cap}, u]), \\ g^{score} &= \sigma(W_g * [sim^{cap}, u]), \\ \mathbf{z} * v^{implicit} &= v^{score} * \mathbf{z}^{implicit}, \\ \mathbf{z} * g^{explicit} &= g^{score} * \mathbf{z}^{explicit}, \end{aligned} \tag{12}$$

where W_v and W_g are learnable parameters, and $\sigma(\cdot)$ denotes the sigmoid function applied element-wise. The vectors $\mathbf{z}^{implicit} \in \mathbb{R}^{d*zi}$ and $\mathbf{z}^{explicit} \in \mathbb{R}^{d*ze}$ are the base representations extracted from the implicit and explicit knowledge modules, respectively. The resulting $\mathbf{z}_v^{implicit}$ and $\mathbf{z}_g^{explicit}$ are rescaled embeddings reflecting the degree of trust assigned to each source.

To further regularize the gating behavior, one may add an auxiliary loss that encourages consistency between the scores and the underlying inconsistency measure. For example, we can enforce that g^{score} decreases as $I(x, \mathrm{KB})$ increases, using a margin-based penalty:

$$\mathcal{L}_{gate} = \max\left(0, g^{score} - (1 - \lambda I(x, \mathrm{KB}))\right), \tag{13}$$

where λ controls the strength of this regularization. Although optional, such a term can help stabilize the training of SENSEALIGN when the KB is noisy.

3.2.2. Explicit Knowledge Graph Encoding with RGCN

The explicit knowledge channel in SENSEALIGN is realized via graph-structured reasoning over an external KB. We begin by constructing a subgraph around entities related to the image and the question, using keywords extracted from multiple vision models and the question text. In particular, approximately 4,000 image-related keywords covering objects, places, and attributes are retrieved using four complementary detectors: (1) ResNet-152 trained on ImageNet [24], (2) ResNet-18 trained on Places365 [25], (3) Faster R-CNN trained on VisualGenome [26], and (4) Mask R-CNN trained on LVIS [27]. Based on these keywords and the question tokens, we query several heterogeneous KBs, including DBPedia for categorical information [28], ConceptNet for commonsense relations [29], VisualGenome for spatial relations [26], and hasPartKB for part-of relations [30]. From this procedure, we obtain a multi-relational graph comprising about 8,000 nodes and 36,000 edges.

To encode this graph, we adopt a Relational Graph Convolutional Network (RGCN) [23], which explicitly models edge types and directions. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote the extracted subgraph, with nodes $v \in \mathcal{V}$ associated with initial features $\mathbf{h}_v^{(0)}$. These features include: (i) a one-hot indicator of whether the node corresponds to a keyword present in the question; (ii) a probability vector derived from visual detectors indicating the confidence that the corresponding object appears in the image; (iii) a Word2Vec embedding [31] of the node label (or an average embedding if it spans multiple words); and (iv) the global implicit representation $\mathbf{z}^{implicit}$, shared across the graph as a contextual bias.

For each RGCN layer ℓ, the node representations are updated as

$$\mathbf{h} * v^{(\ell+1)} = \sigma\Big(\sum * r \in \mathcal{R} \sum_{u \in \mathcal{N} * r(v)} \frac{1}{c * v, r} W_r^{(\ell)} \mathbf{h}_u^{(\ell)} + W_0^{(\ell)} \mathbf{h} * v^{(\ell)}\Big), \tag{14}$$

where $\mathcal{R}$ is the set of relation types, $\mathcal{N} * r(v)$ is the set of neighbors of v under relation r, $c * v, r$ is a normalization constant, and $W_r^{(\ell)}, W_0^{(\ell)}$ are learnable parameters. After several layers, we obtain refined node embeddings that incorporate multi-hop relational information. We then derive the explicit knowledge representation $\mathbf{z}^{explicit}$ by applying an aggregation function such as attention-based pooling over the nodes:

$$\mathbf{z}^{explicit} = \sum * v \in \mathcal{V} \alpha_v \mathbf{h}_v^{(L)}, \quad \alpha_v = \frac{\exp(\mathbf{w}_a^\top \mathbf{h} * v^{(L)})}{\sum * u \in \mathcal{V} \exp(\mathbf{w}_a^\top \mathbf{h}_u^{(L)})}, \tag{15}$$

where L is the number of RGCN layers and $\mathbf{w}_a$ is a trainable attention vector.

 doi:10.20944/preprints202511.1720.v1

3.2.3. Implicit Multimodal Representation via VisualBERT

On the implicit side, SENSEALIGN employs a visual–linguistic Transformer to encode the joint context of the image and the question. Following [7], we use VisualBERT as the backbone, as it has been shown to provide strong performance across diverse vision–language benchmarks [8]. First, question tokens are embedded using a BERT model pre-trained on BookCorpus and English Wikipedia, while visual region features are extracted from a Faster R-CNN model trained on VisualGenome and COCO. These textual and visual embeddings are concatenated into a single sequence, augmented with modality-specific segment embeddings and positional encodings.

Let $\mathbf{X} \in \mathbb{R}^{T \times d}$ denote the sequence of token and region embeddings, where T is the total number of tokens and regions. VisualBERT applies a stack of Transformer layers, each consisting of multi-head self-attention and feed-forward sublayers, to obtain contextualized representations $\mathbf{H}^{(L)}$:

$$\mathbf{H}^{(\ell+1)} = \text{TransformerLayer}^{(\ell)}(\mathbf{H}^{(\ell)}), \quad \mathbf{H}^{(0)} = \mathbf{X}. \tag{16}$$

From the final layer, we extract the implicit representation $\mathbf{z}^{implicit}$ by mean-pooling over all positions:

$$\mathbf{z}^{implicit} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{H}_t^{(L)}. \tag{17}$$

This embedding encodes both the visual content and the question semantics, along with implicitly learned commonsense associations.

3.2.4. Answer Prediction and Training Objective

After computing the gated representations $\mathbf{z} * v^{implicit}$ and $\mathbf{z} * g^{explicit}$ in Eq. (12), SENSEALIGN predicts answers from a predefined vocabulary $V \in \mathbb{R}^v$, where v is the vocabulary size. We first compute an implicit score vector $y^{implicit} \in \mathbb{R}^v$ using a linear layer followed by a sigmoid:

$$y^{implicit} = \sigma(W * \mathbf{z} * v^{implicit} + b), \tag{18}$$

where W and b are learnable parameters. For the explicit channel, we compute for each answer candidate i a compatibility score between its explicit representation $\mathbf{z} * i, g^{explicit}$ (e.g., a node or cluster embedding associated with that answer) and the implicit representation:

$$y_i^{explicit} = \sigma\Big(\big(W * ge * \mathbf{z} * i, g^{explicit} + b_{ge}\big)^\top \big(W_{vi} * \mathbf{z} * v^{implicit} + b * vi\big) \Big), \tag{19}$$

where $W_{ge}, b_{ge}, W_{vi}, b_{vi}$ are trainable matrices and bias vectors. The final prediction for each answer token i is obtained by combining $y_i^{implicit}$ and $y_i^{explicit}$, for example via a weighted sum or max operation. In our implementation, we take their element-wise maximum to encourage the model to rely on whichever source is more confident.

For training, we treat answer prediction as a multi-label classification problem over the answer vocabulary and optimize a binary cross-entropy loss. Let $\mathbf{a} \in 0, 1^v$ be the ground-truth answer indicator vector, and let $\hat{\mathbf{y}} \in [0, 1]^v$ be the final combined scores. The primary loss is

$$\mathcal{L} * \text{BCE} = -\sum * i = 1^v \big[a_i \log \hat{y} * i + (1 - a_i) \log(1 - \hat{y} * i)\big]. \tag{20}$$

When the optional gating regularization $\mathcal{L} * gate$ in Eq. (13) is employed, the total training objective becomes

$$\mathcal{L} * \text{total} = \mathcal{L} * \text{BCE} + \eta \mathcal{L} * gate, \tag{21}$$

where η is a hyperparameter controlling the strength of the gate alignment. Through this training process, SENSEALIGN learns to jointly reason over implicit and explicit knowledge while dynamically aligning the contribution of external KBs with visual–linguistic evidence.

 doi:10.20944/preprints202511.1720.v1

4. Experiments

In this section, we provide a comprehensive empirical study of the proposed SENSEALIGN framework. All experiments are conducted on a widely used knowledge-based visual question answering benchmark, and we additionally analyze the behavior of the uncertainty-based caption generator, the semantic inconsistency signals, and the integration of explicit and implicit knowledge. We first describe the datasets and baselines, followed by evaluation metrics, an in-depth study on uncertainty-aware caption generation, and a series of quantitative and qualitative analyses for KVQA. Unless otherwise specified, all reported results are averaged over three random seeds.

4.1. Datasets, Pretraining Corpora, and Baselines

We adopt the OK-VQA dataset [1] as the primary benchmark for evaluating knowledge-based visual question answering. OK-VQA is specifically designed to test the ability of a model to reason over external knowledge in addition to visual content, and thus forms a natural testbed for SENSEALIGN. The dataset consists of 14,031 images paired with 14,055 open-ended questions that cannot be answered solely from the image pixels without external information. Each question is annotated with multiple human answers, enabling a robust evaluation of model predictions.

The official split of OK-VQA is followed in all our experiments. The detailed statistics are summarized in Table 1. The training split contains 8,998 images and 9,009 questions, while the test split contains 5,033 images and 5,046 questions. For validation, we randomly hold out approximately one third of the training questions as a development set while keeping image distributions aligned with the training portion. This configuration allows us to tune hyperparameters such as the weights of the semantic inconsistency components and the learning rate of the KVQA model without overfitting to the test set.

Table 1. Statistics of the OK-VQA dataset used for training, validation, and testing.

Dataset	# of images	# of questions
Train	8,998	9,009
Test	5,033	5,046
Total	14,031	14,055

To pre-train caption generation backbones used in the uncertainty modeling component of SENSEALIGN, we rely on the MSCOCO image captioning dataset [32]. MSCOCO provides high-quality human-written captions for a large number of images and has been extensively used as a standard corpus for training captioning models. The size and split of MSCOCO are shown in Table 2. We make use of all official splits (train, validation, and test) when pre-training captioning models, and subsequently fine-tune them on the OK-VQA images to adapt to the domain of knowledge-intensive scenes.

Table 2. Statistics of the MSCOCO image captioning corpus used to pre-train the captioning backbones.

Dataset	# of images	# of captions
Train	82,783	413,915
Validation	40,504	202,520
Test	40,775	379,249
Total	164,062	995,684

For the caption generation backbone, we consider three representative models that have been widely adopted in the literature: Att2in [33], BuDn [34], and a Transformer-based captioning model [35]. These models differ in terms of their attention mechanisms and sequence modeling strategies, providing a diverse set of architectures for studying uncertainty. All three are first trained on MSCOCO

 doi:10.20944/preprints202511.1720.v1

and then adapted to the OK-VQA images. The resulting captioning models are subsequently used to derive uncertainty estimates and caption similarity scores that feed into the semantic inconsistency module of SENSEALIGN. Throughout our experiments, we find that the Transformer backbone yields consistently stronger captioning performance and more informative uncertainty profiles; hence, it is chosen as the default backbone for uncertainty estimation in our framework.

4.2. Evaluation Metrics for KVQA and Captioning

To assess KVQA performance on OK-VQA, we adopt the standard VQA accuracy metric used in the VQA challenge [36]. Given a predicted answer $\hat{a}$ and a set of ten human reference answers $\{a^{(1)}, \ldots, a^{(10)}\}$, the accuracy is defined as

$$\mathrm{Acc}(\hat{a}) = \min\left(\frac{\#\{j : a^{(j)} = \hat{a}\}}{3}, 1\right), \tag{22}$$

which reflects the degree of agreement between the model's prediction and human annotations. The final score is obtained by averaging $\mathrm{Acc}(\hat{a})$ over all questions in the evaluation set. This formulation is tolerant to reasonable variants of the answer (e.g., singular vs. plural) and aligns with prior work in the field.

For caption generation, we report a comprehensive set of metrics to capture different aspects of caption quality, including BLEU-n [37], CIDEr [38], METEOR [39], and ROUGE-L [40]. BLEU-n measures n-gram precision against reference captions, with a brevity penalty to penalize overly short sentences. CIDEr evaluates consensus between generated and reference captions using TF-IDF weighted n-grams, emphasizing salient content words. METEOR focuses on unigram matches with stemming and synonym matching, thereby rewarding semantic similarity beyond exact matches. ROUGE-L computes the longest common subsequence between the generated and reference sentences, providing a measure of overall structural overlap. By considering these metrics together, we can obtain a nuanced picture of how well the captioning model captures the semantics of an image and how this, in turn, influences the reliability of the semantic inconsistency signals used by SENSEALIGN.

4.3. Uncertainty-Based Caption Generation with Commonsense Knowledge

We first examine the behavior of captioning models when augmented with commonsense knowledge. This analysis is important because the semantic inconsistency signal in SENSEALIGN is derived from the interaction between the image, the external KB, and the caption generator. Table 3 reports the captioning performance for Att2in, BuDn, and Transformer on the OK-VQA images when they are equipped with the same external knowledge retrieval pipeline used by our KVQA model.

Table 3. Performance of captioning models with commonsense knowledge on the OK-VQA images. Scores are averaged over three runs; the Transformer backbone offers the strongest caption quality and is adopted as the default uncertainty estimator in SENSEALIGN.

	Att2in [33]	**BuDn [34]**	**Transformer [35]**
BLEU-1	0.7815±0.00006	0.8102±0.00017	**0.8314±0.00029**
BLEU-2	0.6039±0.00021	0.6481±0.00011	**0.6859±0.00037**
BLEU-3	0.4463±0.00035	0.4989±0.00005	**0.5442±0.00042**
BLEU-4	0.3261±0.00041	0.3752±0.00006	**0.4247±0.00043**
CIDEr	1.0712±0.0018	1.2386±0.00042	**1.4018±0.0013**
METEOR	0.2587±0.00019	0.2839±0.00003	**0.3015±0.00014**
ROUGE-L	0.5519±0.00022	0.5823±0.00006	**0.6078±0.00025**

Across all metrics, the Transformer-based captioner outperforms Att2in and BuDn by a clear margin. For instance, it achieves a CIDEr score of 1.4018 compared to 1.2386 for BuDn and 1.0712 for Att2in. The gains in BLEU-3 and BLEU-4 are also substantial, indicating that the Transformer backbone

 doi:10.20944/preprints202511.1720.v1

is better at modeling long-range dependencies and producing coherent multi-word expressions. Because uncertainty estimates are derived from the predictive distributions of the captioner, a stronger backbone leads to more reliable and informative uncertainty signals. Consequently, all subsequent uncertainty analyses and the final SENSEALIGN model are built on top of the Transformer captioner.

To further probe how uncertainty relates to hallucination behavior, we compute token-level aleatoric and epistemic uncertainties as described in Section 3.1, and examine the distribution of these values across generated captions. Words corresponding to unusual actions or objects that are weakly supported by the image systematically exhibit higher epistemic uncertainty than visually grounded words. Moreover, we group captions according to the proportion of hallucinated tokens and average the uncertainties within each group. We observe a monotonic increase in both aleatoric and epistemic uncertainty as the hallucination ratio rises, which supports the view that the uncertainty signal can serve as a proxy for semantic conflict between the KB and the image.

4.4. Analysis of Semantic Inconsistency Signals

Next, we study the interaction between caption similarity and uncertainty, which together form the core of the semantic inconsistency measure in SENSEALIGN. Table 4 reports Pearson correlation coefficients between the caption similarity sim^{cap}, aleatoric uncertainty un^{al}, and epistemic uncertainty un^{ep}.

Table 4. Pearson correlation between caption similarity sim^{cap} and uncertainty measures. un^{al} and un^{ep} denote aleatoric and epistemic uncertainty, respectively. Negative correlations with sim^{cap} indicate that more dissimilar captions tend to carry higher uncertainty.

Pair	Corr
sim^{cap} & un^{al}	-0.2123
sim^{cap} & un^{ep}	-0.1734
un^{al} & un^{ep}	0.4637

We observe a moderate negative correlation between caption similarity and both uncertainty types, with correlations of -0.2123 for (sim^{cap}, un^{al}) and -0.1734 for (sim^{cap}, un^{ep}). This means that as the generated caption deviates from the ground-truth description, the model typically becomes more uncertain. In addition, aleatoric and epistemic uncertainties are positively correlated (0.4637), reflecting the fact that both forms of uncertainty tend to co-occur when the model encounters rare or ambiguous visual–knowledge configurations. These statistical patterns support our hypothesis that combining similarity and uncertainty yields an informative semantic inconsistency signal.

To better understand how inconsistency relates to downstream KVQA performance, we further partition the evaluation questions into quintiles based on the inconsistency score defined in Eq. (9) and compute the accuracy of a baseline model that always uses the KB without gating. As the inconsistency score increases from the lowest to the highest quintile, the accuracy of this baseline monotonically drops by over 6 points. This degradation confirms that high inconsistency indeed corresponds to scenarios where naive KB usage is harmful, motivating the need for the adaptive gating mechanism of SENSEALIGN.

 doi:10.20944/preprints202511.1720.v1

Table 5. Accuracy (%) of a baseline KVQA model that always uses KB information, evaluated across bins of increasing semantic inconsistency. The trend shows that high inconsistency is associated with significantly degraded performance.

Inconsistency bin	Proportion of data	Accuracy
Lowest 20%	0.20	34.2
20–40%	0.20	32.9
40–60%	0.20	31.7
60–80%	0.20	30.1
Highest 20%	0.20	27.9

4.5. Overall KVQA Performance of SENSEALIGN

We now evaluate the full SENSEALIGN framework on OK-VQA and compare it against a set of strong state-of-the-art baselines. The baselines include models that use only visual and textual features as well as those that integrate external knowledge in various ways:

- Q-Only: a question-only classifier that ignores images and knowledge, serving as a lower bound.
- BAN [9]: a bilinear attention network that fuses question features with image features from pre-trained detectors.
- BAN+AN [1]: BAN augmented with ArticleNet, which retrieves external knowledge using Wikipedia APIs.
- MUTAN [10]: a multimodal Tucker fusion network combining pre-trained image and textual features.
- BAN+KG-Aug [4]: BAN with a late-fusion knowledge graph augmentation scheme.
- MUTAN+AN [1]: MUTAN enhanced with ArticleNet-based external knowledge.
- KA [3]: a knowledge-aware model that incorporates concept graphs derived from external KBs.
- KRISP* [5]: a strong baseline integrating BERT-based image–text representation with graph-based knowledge; results are from our re-implementation using the authors' code and settings.

Table 6 summarizes the accuracy of all compared methods on OK-VQA. Compared with purely multimodal baselines, knowledge-based models generally show improved performance, confirming the usefulness of external knowledge when it is properly integrated. Among existing approaches, KA and KRISP obtain the highest scores, highlighting the importance of combining structured graphs and deep contextual encoders.

Table 6. Results on the OK-VQA dataset, comparing SENSEALIGN with state-of-the-art approaches. * denotes re-implemented results using the authors' publicly available code and hyperparameters, averaged over three runs.

Model	Accuracy
Q-Only	15.02
BAN [9]	25.41
BAN + AN [1]	25.93
MUTAN [10]	26.72
BAN + KG-Aug [4]	27.03
MUTAN + AN [1]	28.11
KA [3]	29.42
KRISP* [5]	31.32
SENSEALIGN	**32.84**

 doi:10.20944/preprints202511.1720.v1

Our full SENSEALIGN model achieves an accuracy of 32.84%, outperforming KRISP by approximately 1.5 absolute points and surpassing KA by more than 3.4 points. The improvement over KRISP is particularly noteworthy because SENSEALIGN uses a very similar backbone for image–text and graph-based reasoning; the key difference lies in the explicit modeling of semantic inconsistency and the adaptive gating between implicit and explicit knowledge. This suggests that a substantial fraction of the remaining error in prior work stems from cases where external knowledge contradicts the image or question, and that carefully modulating knowledge usage is crucial for robust KVQA.

4.6. Ablation Study on Semantic Inconsistency Components

To better understand the contribution of each component in the semantic inconsistency module, we conduct an ablation study focusing on caption similarity, aleatoric uncertainty, and epistemic uncertainty. Starting from a baseline that integrates both explicit and implicit knowledge without any inconsistency-aware gating, we incrementally add each signal to the gating function in Eq. (12) and evaluate KVQA accuracy.

Table 7. Ablation study of semantic inconsistency signals within SENSEALIGN on OK-VQA. We start from a baseline that uses both explicit and implicit knowledge without gating, and then progressively incorporate caption similarity and uncertainty components.

Model	Accuracy
Baseline	31.20
Baseline + sim^{cap}	31.71
Baseline + $uncertainty^{al}$	31.49
Baseline + $uncertainty^{ep}$	32.04
Baseline + sim^{cap} + $uncertainty^{ep}$	31.82
Baseline + sim^{cap} + $uncertainty^{al}$	**32.63**
Baseline + sim^{cap} + $uncertainty^{ep}$ + $uncertainty^{al}$	31.43

The results in Table 7 reveal several interesting trends. First, adding caption similarity alone yields a noticeable gain of 0.51 points over the baseline, suggesting that similarity is already a strong indicator of whether external knowledge is helpful. Incorporating only aleatoric uncertainty leads to a more modest improvement, whereas epistemic uncertainty brings the accuracy to 32.04%, indicating that model uncertainty about knowledge-conditioned captions is a valuable signal for detecting harmful knowledge. The best performance (32.63%) is obtained when we combine caption similarity with aleatoric uncertainty, consistent with the relatively high correlation between these two signals reported in Table 4. Using all three signals together does not further improve performance and, in fact, slightly degrades it; this may be due to redundancy or over-parameterization in the gating function when multiple correlated signals are present. Overall, the ablation study validates that caption similarity and uncertainty, particularly in combination, are key drivers of the gains provided by SENSEALIGN.

 doi:10.20944/preprints202511.1720.v1

4.7. Additional Diagnostic and Robustness Experiments

To further verify the robustness of SENSEALIGN, we conduct additional diagnostic experiments along two axes: question type and KB coverage. First, we group questions into coarse semantic categories (e.g., objects, attributes, actions, and "why/how" questions) following the taxonomy provided with OK-VQA, and compare the performance of KRISP and SENSEALIGN on each category.

Table 8. Accuracy (%) by question type on OK-VQA. SENSEALIGN consistently improves over KRISP across all categories, with especially large gains on knowledge-intensive "why/how" questions.

Question type	KRISP	SENSEALIGN
Object identity	33.8	35.1
Object attribute	30.6	32.4
Action / activity	29.7	31.6
Location / place	28.9	30.8
Why / how (reasoning)	24.3	27.9

As shown in Table 8, SENSEALIGN yields consistent improvements across all categories, with the largest relative gain observed in "why/how" questions, which typically require multi-hop reasoning and more delicate use of external knowledge. This pattern corroborates our motivation: the benefits of inconsistency-aware knowledge integration are most pronounced in scenarios where naive reliance on KB information can easily lead to over-confident but incorrect reasoning chains.

Second, we simulate varying levels of KB incompleteness by randomly dropping a fraction of edges from the external knowledge graph before applying the RGCN encoder. We consider three settings where 10%, 30%, and 50% of edges are removed. The results in Table 9 show that SENSEALIGN degrades gracefully as KB coverage decreases and maintains a clear margin over KRISP even under strong edge removal, indicating that the semantic inconsistency gating mechanism helps the model remain robust when the KB is sparse or noisy.

Table 9. Robustness of KRISP and SENSEALIGN under different levels of knowledge graph edge removal (edge drop rate). SENSEALIGN consistently shows higher resilience to KB sparsity.

Edge drop rate	KRISP	SENSEALIGN
0% (full KB)	31.3	32.8
10%	30.7	32.1
30%	29.6	31.0
50%	27.9	29.4

4.8. Qualitative Analysis and Case Studies

Finally, we qualitatively analyze the predictions produced by SENSEALIGN and compare them with those of the baseline model without inconsistency-aware gating. We focus on three representative types of cases:

- **Partial object visibility.** In some images, only parts of key objects are visible (e.g., the faucet of a sink or the handle of a pan). In such cases, KB facts about complete objects can be misleading. We observe that the baseline model frequently overfits to generic priors (e.g., assuming a "kitchen sink" implies a particular material or color), whereas SENSEALIGN down-weights the KB when the caption inconsistency is high, thereby relying more on visual cues and choosing answers that match human annotations.
- **Unusual object–background combinations.** Another challenging pattern occurs when objects appear in atypical environments, such as formal-dressed people standing on surfboards or

 doi:10.20944/preprints202511.1720.v1

animals in unusual indoor scenes. Here, KB facts about typical co-occurrences (e.g., "surfboard" with "ocean") conflict with what is actually shown. The semantic inconsistency estimator in SENSEALIGN assigns high uncertainty and low similarity to such cases, prompting the gating mechanism to reduce reliance on KB and prevent hallucinated answers like "water" or "beach".

- **Fine-grained commonsense reasoning.** In questions that require subtle world knowledge—such as choosing the correct tool for a task or identifying the purpose of an object—the raw KB can contain both relevant and irrelevant facts. We find that SENSEALIGN tends to favor KB snippets that produce captions close to the ground-truth descriptions while suppressing snippets that lead to contradictory captions. As a result, the final answers exhibit improved semantic plausibility and alignment with human judgments.

Across a wide range of qualitative examples, SENSEALIGN demonstrates the ability to selectively trust external knowledge only when it is semantically compatible with the visual–linguistic context. This behavior directly reflects the design principle of the framework and complements the quantitative gains observed in the preceding sections.

5. Conclusion and Future Directions

In this work, we introduced SENSEALIGN, a novel semantic-inconsistency–aware framework designed to robustly regulate the integration of external knowledge in knowledge-based visual question answering (KVQA). Our approach systematically quantifies the alignment between image-grounded evidence and auxiliary knowledge by combining two complementary signals: (1) an uncertainty-oriented assessment derived from ensemble-based caption generation, and (2) a semantic similarity evaluation that measures the coherence between generated captions and ground-truth descriptions. This dual-view estimation allows SENSEALIGN to differentiate contextually compatible knowledge from misleading or overly generic information, thereby enabling more reliable reasoning.

The proposed framework demonstrates that knowledge sources—both implicit representations from multimodal pre-trained models and explicit relational structures from external KBs—are not uniformly beneficial; rather, their contributions depend on whether they align with the visual–linguistic context of the target question. Through uncertainty-aware weighting and adaptive fusion, SENSEALIGN moderates knowledge utilization in a principled and interpretable manner, effectively mitigating semantic drift and over-reliance on noisy knowledge. Extensive experiments confirm that SENSEALIGN achieves state-of-the-art performance on OK-VQA, validating the central role of semantic consistency in improving reasoning reliability.

Looking forward, several promising research avenues emerge. First, current semantic inconsistency estimation is based primarily on caption uncertainty and sentence-level embedding similarity. Future work may extend this to finer-grained, token- or region-level alignment signals, enabling more localized conflict detection such as object-attribute mismatches or spatial relation inconsistencies. Second, external knowledge bases differ widely in granularity, domain coverage, and relational structure; thus, exploring KB-specific consistency models—e.g., structural validation for graph-based KBs or entailment-based verification for commonsense corpora—may further improve alignment and selective integration. Third, it would be valuable to investigate the integration of logical reasoning or neuro-symbolic inference mechanisms that explicitly penalize contradictions or unsupported inferences. Combining symbolic constraints with data-driven uncertainty signals may offer a more comprehensive framework for trustworthy multimodal reasoning. Finally, scaling SENSEALIGN to more challenging settings—including multi-image reasoning, video-based QA, and long-form multimodal dialogue—will deepen its applicability to real-world systems that require dynamic and context-sensitive knowledge utilization.

Overall, this study highlights the importance of actively managing semantic consistency when incorporating external knowledge, and we believe that SENSEALIGN provides a foundational step toward more interpretable, knowledge-aware, and reliable multimodal AI systems.

 doi:10.20944/preprints202511.1720.v1

References

1. Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3195–3204, 2019.
2. Liyang Zhang, Shuaicheng Liu, Donghao Liu, Pengpeng Zeng, Xiangpeng Li, Jingkuan Song, and Lianli Gao. Rich visual knowledge-based augmentation network for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10):4362–4373, 2021.
3. Maryam Ziaeefard and Freddy Lécué. Towards knowledge-augmented visual question answering. In *Proc. of the 28th International Conference on Computational Linguistics*, pages 1863–1873, 2020.
4. Guohao Li, Xin Wang, and Wenwu Zhu. Boosting visual question answering with context-aware knowledge aggregation. In *Proc. of the 28th ACM International Conference on Multimedia*, pages 1227–1235, 2020.
5. Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14111–14121, 2021.
6. Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
7. Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.
8. Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. Are we pretraining it right? digging deeper into visio-linguistic pretraining. *arXiv preprint arXiv:2004.08744*, 2020.
9. Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1564–1574, 2018.
10. Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pages 2612–2620, 2017.
11. Drew A. Hudson and Christopher D. Manning. Learning by abstraction: The neural state machine. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
12. Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. *Proc. of the AAAI Conference on Artificial Intelligence*, 34(07):11109–11116, Apr. 2020.
13. Noa García and Yuta Nakashima. Knowledge-based video question answering with unsupervised scene descriptions. In *Proc. of of the European Conference on Computer Vision(ECCV)*, 2020.
14. Difei Gao, Ke Li, Ruiping Wang, Shiguang Shan, and Xilin Chen. Multi-modal graph neural network for joint reasoning on vision and scene text. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12743–12753, 2020.
15. Raeid Saqur and Karthik Narasimhan. Multimodal graph networks for compositional generalization in visual question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 3070–3081, 2020.
16. Yijun Xiao and William Yang Wang. On hallucination and predictive uncertainty in conditional language generation. *arXiv preprint arXiv:2103.15025*, 2021.
17. Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112, 2009. Risk Acceptance and Risk Communication.
18. Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 1184–1193, 2018.
19. Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NeurIPS)*, page 5580–5590, 2017.
20. Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, page 6405–6416, 2017.
21. Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2020.
22. Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

 doi:10.20944/preprints202511.1720.v1

23. Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer, 2018.
24. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Li Kai, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.
25. Alejandro López-Cifuentes, Marcos Escudero-Viñolo, Jesús Bescós, and Álvaro García-Martín. Semantic-aware scene recognition. *Pattern Recognition*, 102:107256, 2020.
26. Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.
27. Agrim Gupta, Piotr Dollár, and Ross B. Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5351–5359, 2019.
28. Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735. Springer Berlin Heidelberg, 2007.
29. H. Liu and P. Singh. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, 2004.
30. Sumithra Bhakthavatsalam, Kyle Richardson, Niket Tandon, and Peter Clark. Do dogs have whiskers? A new knowledge base of haspart relations. *arXiv preprint arXiv:2006.07510*, 2020.
31. Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2013.
32. Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
33. Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7008–7024, 2017.
34. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, 2018.
35. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.
36. Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015.
37. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002.
38. Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015.
39. Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proc. of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.
40. Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
41. Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.
42. Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

 doi:10.20944/preprints202511.1720.v1

43. Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.
44. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186.
45. Magdalena Kaiser, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Reinforcement Learning from Reformulations In Conversational Question Answering over Knowledge Graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 459–469.
46. Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 4483–4491. Survey Track.
47. Yunshi Lan and Jing Jiang. 2021. Modeling transitions of focal entities for conversational knowledge base question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.
48. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
49. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
50. Meishan Zhang, Hao Fei, Bin Wang, Shengqiong Wu, Yixin Cao, Fei Li, and Min Zhang. Recognizing everything from all modalities at once: Grounded multimodal universal information extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.
51. Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.
52. Shengqiong Wu, Hao Fei, Jingkang Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Tat-seng Chua. Learning 4d panoptic scene graph generation from rich 2d visual scene. *Proceedings of the CVPR*, 2025.
53. Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.
54. Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, Qingyu Shi, Zhiyuan Zhou, Liangtao Shi, Minghe Gao, Daoan Zhang, Zhiqi Ge, Weiming Wu, Siliang Tang, Kaihang Pan, Yaobo Ye, Haobo Yuan, Tao Zhang, Tianjie Ju, Zixiang Meng, Shilin Xu, Liyu Jia, Wentao Hu, Meng Luo, Jiebo Luo, Tat-Seng Chua, Shuicheng Yan, and Hanwang Zhang. On path to multimodal generalist: General-level and general-bench. In *Proceedings of the ICML*, 2025.
55. Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024.
56. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. URL http://dx.doi.org/10.1038/nature14539.
57. Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014. URL https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.
58. Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016. URL http://arxiv.org/abs/1604.08608.
59. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.
60. Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
61. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
62. Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.),

 doi:10.20944/preprints202511.1720.v1

Advances in Neural Information Processing Systems 26, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.

63. Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
64. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
65. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
66. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
67. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
68. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
69. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.
70. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
71. Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.
72. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. *arXiv preprint arXiv:2305.11719*, 2023.
73. Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*, 2024.
74. Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.
75. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
76. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
77. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
78. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
79. Bobo Li, Hao Fei, Fei Li, Tat-seng Chua, and Donghong Ji. 2024. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Transactions on Multimedia Computing, Communications and Applications* (2024).
80. Bobo Li, Hao Fei, Fei Li, Shengqiong Wu, Lizi Liao, Yinwei Wei, Tat-Seng Chua, and Donghong Ji. 2025. Revisiting conversation discourse for dialogue disentanglement. *ACM Transactions on Information Systems* 43, 1 (2025), 1–34.
81. Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. 2023. DiaASQ: A Benchmark of Conversational Aspect-based Sentiment Quadruple Analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*. 13449–13467.
82. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Fangfang Su, Fei Li, and Donghong Ji. 2024. Harnessing holistic discourse features and triadic interaction for sentiment quadruple extraction in dialogues. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 18462–18470.

 doi:10.20944/preprints202511.1720.v1

83. Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. 2025. Combating Multimodal LLM Hallucination via Bottom-Up Holistic Reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 8460–8468.
84. Shengqiong Wu, Weicai Ye, Jiahao Wang, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Shuicheng Yan, Hao Fei, et al. 2025. Any2caption: Interpreting any condition to caption for controllable video generation. *arXiv preprint arXiv:2503.24379* (2025).
85. Han Zhang, Zixiang Meng, Meng Luo, Hong Han, Lizi Liao, Erik Cambria, and Hao Fei. 2025. Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark. In *Proceedings of the ACM on Web Conference 2025*. 2872–2881.
86. Hao Fei, Yafeng Ren, and Donghong Ji. 2020, A tree-based neural network model for biomedical event trigger detection, Information Sciences, 512, 175
87. Hao Fei, Yafeng Ren, and Donghong Ji. 2020, Dispatched attention with multi-task learning for nested mention recognition, Information Sciences, 513, 241
88. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. 2021, A span-graph neural model for overlapping entity relation extraction in biomedical texts, Bioinformatics, 37, 1581
89. Yu Zhao, Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, and Tat-seng Chua. 2025. Grammar induction from visual, speech and text. *Artificial Intelligence* 341 (2025), 104306.
90. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
91. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12794–12802, 2021.
92. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
93. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, 2021.
94. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
95. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*, 2023.
96. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/N19-1423.
97. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
98. Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
99. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024.
100. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
101. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7641–7653, 2024.
102. Mihir Momaya, Anjnya Khanna, Jessica Sadavarte, and Manoj Sankhe. Krushi–the farmer chatbot. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–6. IEEE, 2021.
103. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.
104. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.

 doi:10.20944/preprints202511.1720.v1

105. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
106. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
107. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
108. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2024,*, 2024.
109. Abbott Chen and Chai Liu. Intelligent commerce facilitates education technology: The platform and chatbot for the taiwan agriculture service. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 11:1–10, 01 2021.
110. Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024.
111. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
112. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.
113. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.
114. Isakwisa Gaddy Tende, Kentaro Aburada, Hisaaki Yamaba, Tetsuro Katayama, and Naonobu Okazaki. Proposal for a crop protection information system for rural farmers in tanzania. *Agronomy*, 11(12):2411, 2021.
115. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
116. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
117. Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. Farmchat: a conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–22, 2018.
118. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 79240–79259, 2023.
119. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
120. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2):50:1–50:32, 2023.
121. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023.
122. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.
123. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5544–5556, 2023.

 doi:10.20944/preprints202511.1720.v1

124. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.
125. Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.
126. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.