Crafting Descriptive Information for a Zero-shot Method to Improve Knowledge-Based Visual Question Answering Performance

Mohammad Mahdi Moradi, Sudhir Mudur
1Concordia University, Montreal, Canada

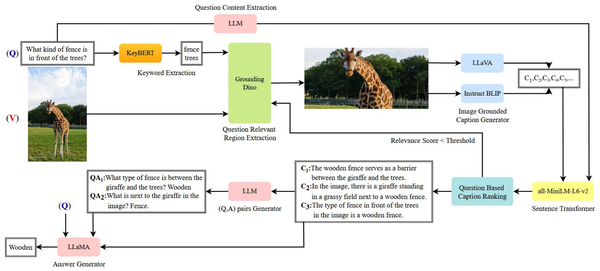
WACV 2026 TUCSON, AZ 3/6 – 3/10

Abstract

Knowledge-based visual question answering (KB-VQA) requires integrating visual perception with external knowledge to answer questions that go beyond what is directly observable in an image. While recent zero-shot approaches leverage LLMs as implicit knowledge sources, their performance is often limited by imprecise visual grounding, noisy or redundant auxiliary descriptions, and hallucinated content that distracts the reasoning process.

We propose GC-KBVQA, a modular, training-free, zero-shot framework designed to craft faithful, question-relevant descriptive information that effectively guides LLM-based reasoning. Our approach combines keyword-guided visual grounding to localize question-relevant regions, dual caption generation to capture complementary visual semantics, and semantic filtering with inter-stage feedback to progressively refine auxiliary text while suppressing irrelevant or hallucinated descriptions.

Despite its lightweight design and complete absence of task-specific training or in-context examples, GC-KBVQA achieves state-of-the-art zero-shot performance, outperforming strong baselines by up to +10.97% on OK-VQA, A-OKVQA, and VQAv2. The framework is model-agnostic and scalable, maintaining robust performance across LLMs ranging from TinyLLaMA-1B to LLaMA-3-8B with minimal degradation. Extensive ablation studies further demonstrate that grounding accuracy,



Methods and Materials

GC-KBVQA is a training-free, zero-shot framework for knowledge-based visual question answering. Instead of relying on external knowledge retrieval or fine-tuning, the method focuses on crafting reliable, question-aligned descriptive information to guide LLMs. The framework operates in four lightweight stages, each designed to reduce noise, improve grounding, and suppress hallucination.

1) **Keyword-Guided Visual Grounding:** Key concepts are extracted from the question to guide visual grounding, localizing question-relevant image regions. Counting questions are handled via direct object counting to reduce unnecessary reasoning.
2) **Dual Caption Generation with Semantic Filtering:** Two complementary vision–language models generate diverse captions for grounded regions. Captions are ranked by semantic relevance to the question, retaining only the most informative ones. If too few captions meet the relevance threshold, an inter-stage feedback mechanism re-triggers grounding to recover useful visual evidence.
3) **Caption-Driven QA Pair Construction:** Filtered captions are converted into auxiliary QA pairs that provide lightweight reasoning guidance, helping the LLM focus on task-relevant inference.
4) **Structured Answer Generation:** Captions and QA pairs guide structured zero-shot answer prediction.

Materials and Evaluation Setup: GC-KBVQA is evaluated in a **pure zero-shot setting** on OK-VQA, A-OKVQA, and VQAv2. All components are modular and replaceable, allowing the framework to scale across different LLM sizes and architectures.

Algorithm 1 Compact Caption Generation & Filtering

Require: Image I, Question q, Regions $\{r_j, w_j\}$ where w_j is each region confidence score, VLMs LLaVA/InstructBlip with v and b captions, threshold $\theta = 0.5$, $k = 3$, $T = 2$, $\alpha = 0.8$
Ensure: Captions S^* of size k

$\tilde{q} \leftarrow \text{LLM}(\texttt{Determine the main idea of this question in short:} \| q \|)$
$\mathbf{v}_q \leftarrow \text{Embed}(\tilde{q})$
Pick $r^* = \arg\max_j w_j$, $t=0$, $\tau=0.25$
while $t \leq T$ **do**
  Generate $\{c_i^L\}_{i=1}^{v}$, $\{c_i^B\}_{i=1}^{b}$ on r^*; $C \leftarrow \{c^L, c^B\}$
  for $c \in C$ **do** $s(c) \leftarrow \cos(\mathbf{v}_q, \text{Embed}(c))$
  end for
  $S^* \leftarrow \text{Top-k}\{c : s(c) \geq \theta\}$
  if $|S^*| \geq k$ **then break**
  end if
  $t \leftarrow t + 1$
  if unused r_j with $w_j \geq \tau$ **then** $r^* \leftarrow \arg\max_{j \neq *} w_j$
  else $\tau \leftarrow \alpha\tau$; re-detect; $r^* \leftarrow \arg\max_j w_j$
  end if
end while
if $|S^*| < k$ **then**
  Generate on I; compute $s(c)$; $S^* \leftarrow \text{Top-k}(C)$
end if
return S^*

Results

GC-KBVQA achieves state-of-the-art zero-shot performance on OK-VQA, A-OKVQA, and VQAv2, outperforming strong baselines by up to +10.97% without any task-specific training. The framework remains robust across LLM sizes from 1B to 8B parameters and significantly reduces hallucinated and irrelevant auxiliary content through semantic filtering and feedback.

Method	Size	OK-VQA	A-OKVQA	VQAv2
Zero-shot evaluation without extra end-to-end training				
PICa [42]	175B	17.7	-	-
PNP-VQA [38]	11B	35.9	-	64.8
Img2LLM [16]	30B	41.8	38.7	60.3
Img2LLM [16]	175B	45.6	42.9	61.9
LAMOC [11]	11B	40.3	37.9	-
KGenVQA [8]	11B	45.4	39.1	-
ZVQAF [26]	11B	40.5	37.1	64.3
LLM_Guided_VQA[31]	-	43.7	42.1	61.2
RQPrompt [23]	175B	46.4	43.9	-
DecomVQA [21]	13B	39.79	53.36	-
DIETCOKE [25]	7B	49.2	48.6	-
ours	39B	**54.57**	**53.87**	**67.96**
Zero-shot evaluation with extra end-to-end training				
VLKD [9]	408B	13.3	-	44.5
Flamingo [1]	80B	50.6	-	56.3
BLIP-2 [24]	12B	45.9	-	65.2
Few-shot evaluation				
PICa [42]	175B	46.5	-	54.3
PromptCAP [18]	175B	60.4	56.3	-
Prophet [36]	-	61.1	58.2	-

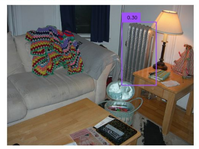
Model	Size	OK-VQA	A-OKVQA	VQAv2
TinyLlaMA-1B	18B	49.45	48.04	66.26
Qwen1.5-1.8B	20B	47.74	45.86	62.79
Mistral-3B	24B	50.30	50.23	67.11
Mistral-7B	36B	52.17	50.95	**68.81**
Llama3-8B	39B	**54.57**	**53.87**	67.96

Variant	CHAIR$_i$ ↓	CHAIR$_s$ ↓
Raw (no filt./fb)	41%	15%
+Filtering only	37%	13%
+Filtering + Feedback	**23%**	**8%**

Variant	Mean ↑	Top-1 ↑	Coverage@0.5 ↑
Raw (no filt./fb)	0.35	0.42	70%
+Filtering only	0.48	0.55	86%
+Filtering + Feedback	**0.55**	**0.60**	**90%**

Question: What is the metal pipe device in old homes that heats them called?

GT: radiator/heater/furnace



KeyWords: metal pipe device

Question Concept: name of the Heating pipe

Caption1: On the end table, a vintage-style radiator (silver color), appearing to be covered with a light-colored insulating material.
Caption2: A vintage radiator stands guard in the corner, a silent sentinel of warmth in this cozy corner of the home.
Caption3: The metal pipe device in the image is an old-fashioned radiator, which is a type of heating system commonly found in older homes.

QA1: What is the device shown in the image? Radiator

QA2: What type of heating system is commonly found in older homes? Radiator

Predicted Answer: Radiator

Question: How many vehicles have their lights on?

GT: 1

KeyWords: vehicles, lights

Predicted Answer: 1

Discussion

Our results indicate that zero-shot KB-VQA performance is primarily constrained by grounding accuracy and auxiliary information quality, rather than LLM scale or fine-tuning. GC-KBVQA improves reliability by enforcing question-aware visual grounding and progressive filtering of descriptive content, which directly reduces hallucination and improves reasoning fidelity. The consistent performance across LLM sizes shows that structured, image-grounded prompts can effectively substitute for task-specific supervision. Moreover, the modular design isolates performance gains to pipeline components rather than parameter count, suggesting a scalable and extensible direction for zero-shot KB-VQA.

Question: Based on the clothing what year was this picture taken?

GT: 1930/1960/1913/1953

KeyWords: year, clothing

Question Concept: year a picture was taken based on the clothing

Caption1: The year a picture was taken based on the clothing depicted in the image is around the 1920s or 1930s.
Caption2: A moment frozen in time: 1945, as seen through the lens of a vintage camera.
Caption3: The image appears to be a vintage black and white photograph, likely from the early to mid-20th century, judging by the clothing and hairstyles of the individuals in the picture.

QA1: What is the likely time period of the picture? Mid-20th century

QA2: What type of photograph is the image likely to be? Black and white

Predicted Answer: 1945

Conclusions

- GC-KBVQA achieves state-of-the-art KB-VQA performance without fine-tuning or external retrieval by crafting question-aligned descriptive information..
- Question-guided grounding, semantic filtering, and inter-stage feedback significantly suppress noisy and hallucinated auxiliary text.
- Robust performance is maintained across LLMs from 1B–8B parameters, demonstrating model-agnostic scalability.
- Gains stem from pipeline design rather than parameter count, highlighting a practical direction for reliable zero-shot KB-VQA.

Contact

Mohammad Mahdi Moradi
Concordua University
Email: moradimohammadmahdi75@gmail.com

References

1. Yang, Zhengyuan, et al. "An empirical study of gpt-3 for few-shot knowledge-based vqa." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 36. No. 3. 2022.
2. Shao, Zhenwei, et al. "Prompting large language models with answer heuristics for knowledge-based visual question answering." *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. 2023.
3. Guo, Jiaxian, et al. "From images to textual prompts: Zero-shot visual question answering with frozen large language models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023.