Precision Empowers, Excess Distracts: Visual Question Answering With Dynamically Infused Knowledge In Language Models

Manas Jhalani, Annervaz K M and Pushpak Bhattacharyya
Computer Science and Engineering, IIT Bombay
Indian Institute of Science, Bangalore
{manasj, pb}@cse.iitb.ac.in
annervaz@iisc.ac.in

Abstract

In the realm of multimodal tasks, Visual Question Answering (VQA) plays a crucial role by addressing natural language questions grounded in visual content. Knowledge-Based Visual Question Answering (KBVQA) advances this concept by adding external knowledge along with images to respond to questions. We introduce an approach for KBVQA, augmenting the existing vision-language transformer encoder-decoder (OFA) model (Wang et al., 2022). Our main contribution involves enhancing questions by incorporating relevant external knowledge extracted from knowledge graphs, using a *dynamic triple extraction* method. We supply a flexible number of triples from the knowledge graph as context, tailored to meet the requirements for answering the question. Our model, enriched with knowledge, demonstrates an average improvement of **4.75%** in Exact Match Score over the state-of-the-art on **three** different KBVQA datasets. Through experiments and analysis, we demonstrate that furnishing variable triples for each question *improves the reasoning capabilities of the language model* in contrast to supplying a fixed number of triples. This is illustrated even for recent large language models. Additionally, we highlight the model's generalization capability by showcasing its SOTA-beating performance on a small dataset, achieved through straightforward fine-tuning.

1 Introduction

The domain of Knowledge-Based Visual Question Answering (KBVQA) not only utilizes visual information extracted from images, such as object attributes and visual relationships but also integrates supporting facts to facilitate accurate reasoning and answer prediction.

Motivation: Recently, large language models (LLMs) like GPT-4 (OpenAI et al., 2023) have garnered attention for their human-like understanding of both images and language, enabling them to tackle KBVQA questions very effectively. However, these LLMs come with a significant drawback: their immense size (around a trillion parameters) poses challenges for offline usage. Additionally, they struggle with user-centric data, such as questions related to named entities within an image. For instance, consider questions like *Who is the person in the middle of the image?* or *What is the age of the person shown in the image?* In such cases, a model should provide specific answers, such as the person's name or age, rather than generic responses like *man* or *I can't guess the age*. This could also limit the performance of many IoT applications where real-time user-centric data plays a crucial role.

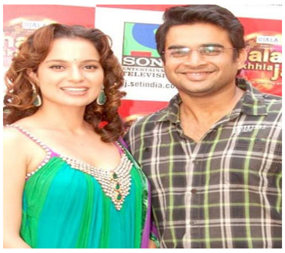


Figure 1: Example question answerable solely from an image (Shah et al., 2019), without requiring external information.
Question: Who is to the right of R.Madhavan?
Named Entities: [Kangana Ranaut, R. Madhavan]

To solve this problem previous works in KBVQA (Li et al., 2020; Garcia-Olano et al., 2021; Vickers et al., 2021) used a fixed number of triples from knowledge graphs as additional sources of information to answer the question. Nevertheless, using a fixed number of triples for all questions may lead to either inadequate information or unnecessary noise, potentially resulting in inaccurate predictions. E.g. in Figure 1, *Who is to the right of R.Madhavan?* These questions can be answered from image features alone and when additional knowledge is given it introduces noise which often confuses the model and subsequently leads to incorrect predictions. Similarly, some questions require more triples to reach the correct answer, but providing a fixed number of triples can limit the model's reasoning capabilities due to insufficient information.

Our Approach: To address this, we propose a **dynamic triple filtering** module capable of retrieving a variable number of triples from knowledge graphs as context to answer the questions. We use an established vision language transformer encoder-decoder (OFA) (Wang et al., 2022) model which takes an image, question, and filtered triples as input to predict the desired answer.

Our contributions are,

1. An approach to Knowledge Based VQA, providing a **dynamic triple filtering** method that gives question-specific triples instead of a fixed number of ones, serving as context to answer the posed question. The approach surpasses the state-of-the-art (SOTA) on three different KBVQA datasets by at least **4.12%**(Section 4, Section 5 & Section 3.1.2).
2. A benchmark across all three datasets. Through a comprehensive evaluation of the VQA model under diverse settings, encompassing both its strengths and weaknesses, we ascertain that the enhanced performance can be attributed to the integration of external knowledge from ConceptNet (Speer et al., 2018) and WikiData (Vrandečić and Krötzsch, 2014) in the form of an additional "knowledge vector". (Section 4)
3. An enhanced knowledge base for the CRIC-VQA dataset. This enhancement raises the number of triples from 3,439, as documented in CRIC-VQA (Gao et al., 2023), to 99,586 triples (Section 4.2)

2 Related Work

Knowledge-based VQA: KBVQA is a recent advancement that incorporates external knowledge along with images and questions to arrive at an answer. There are various datasets published for this purpose. These datasets are mainly of two types-

Open Domain Knowledge-Based VQA involves answering questions that require broad-world knowledge, going beyond what's directly visible in an image. Several datasets, such as OK-VQA(Marino et al., 2019), A-OKVQA(Schwenk et al., 2022), and ScienceQA(Lu et al., 2022), fall into this category. Researchers have tackled this challenge by leveraging various sources of information. Recent work utilizes large language models (LLMs) like GPT-3.5 (Gui et al., 2022; Lin et al., 2022) to retrieve relevant knowledge. Works such as (Khademi et al., 2023; Lin et al., 2022; Gui et al., 2022) found that increasing the diversity of knowledge sources leads to improved accuracy in answering these types of questions.

Closed Domain Knowledge-Based VQA pertains to questions that rely on information from a fixed knowledge base. Datasets like FVQA (Wang et al., 2017), KVQA (Shah et al., 2019), ViQuae (Lerner et al., 2022), and CRIC-VQA (Gao et al., 2023) fall into this category. Some approaches like (Shevchenko et al., 2021; Li et al., 2020), utilize knowledge graphs to retrieve relevant information needed to answer specific questions. Others (Lerner et al., 2024) have employed a fixed multimodal knowledge base, which combines information from different modalities to provide accurate answers.

As user-centric or factual questions require a limited knowledge base to answer a question our work focuses on Closed Knowledge Based VQA. In previous works, the MEMNET architecture (Tai et al., 2017) was utilized. It retrieved relevant facts from knowledge graphs and then passed them to a BI-LSTM (Huang et al., 2015) to find the answer. Recent models have leveraged the Vision+Language BERT model (Su et al., 2020) to obtain desired answers. Another approach, proposed by (Chen et al., 2020), utilizes a BERT-based encoder UNITIER (Devlin et al., 2019) which frames VQA as a classification problem. However, this method has limitations in its applicability to other datasets due to fixed class labels. The latest work, POP-VQA by (Sahu et al., 2024), employs MT-CNN to retrieve a fixed number of highly relevant facts. These relevant facts, along with questions and images, are then fed into a transformer encoder-decoder model to obtain the desired answer.

Datasets: Three primary datasets, KVQA (Shah et al., 2019), CRIC-VQA (Gao et al., 2023), and FVQA (Wang et al., 2017), are used for Closed Domain Knowledge-Based VQA. KVQA contains 183,000 Q&A pairs, emphasizing named entity understanding with 18,000 entities across 24,000 images. Conversely, FVQA and CRIC-VQA prioritize commonsense over named entities, with FVQA having 5826 questions and 2190 images, and CRIC-

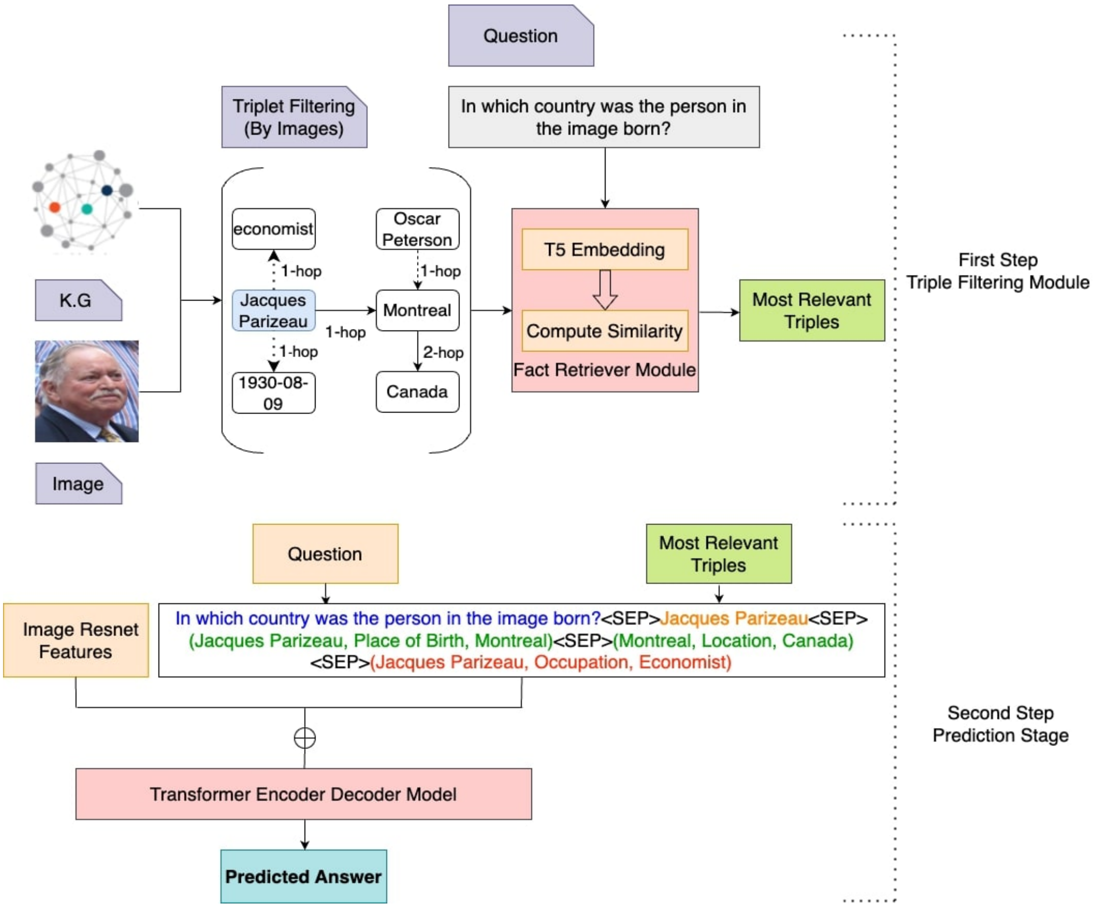

Question
In which country was the person in the image born?
Triplet Filtering (By Images)
K.G
Image
economist
Oscar Peterson
Jacques Parizeau
Montreal
1930-08-09
Canada
1-hop
2-hop
T5 Embedding
Compute Similarity
Fact Retriever Module
Most Relevant Triples
First Step
Triple Filtering Module
Question
Most Relevant Triples
Image Resnet Features
In which country was the person in the image born?<SEP>Jacques Parizeau<SEP>(Jacques Parizeau, Place of Birth, Montreal)<SEP>(Montreal, Location, Canada)<SEP>(Jacques Parizeau, Occupation, Economist)
⊕
Transformer Encoder Decoder Model
Predicted Answer
Second Step
Prediction Stage

Figure 2: The proposed framework is illustrated in the flow diagram. In the first stage of prediction, triples are filtered based on images, followed by an additional round of filtering based on questions. Finally, the extracted triples in green represent useful triples and the triples in red represent noisy ones. In the second stage of prediction, **Relevant Triples**, **Image Resnet Features** and **Questions**, are fed into a transformer encoder-decoder model (OFA) to generate the predicted answer.⊕ represents the concatenation of all the features to pass it to the transformer encoder-decoder to get the predicted answer. Irrelevant triples are depicted with dashed lines, while relevant triples, filtered based on images, are represented with bold lines.

VQA comprising 494K questions and 94K images. We primarily used external knowledge sourced from the Wikidata (Vrandečić and Krötzsch, 2014) and ConceptNet (Speer et al., 2018) knowledge graphs to address questions within these datasets.

3 Our Approach

Our approach follows a two-stage process to determine the answer for a given question. Let $\mathcal{A}$ be the set of potential answers, $\mathcal{I}$ be the set of images, i be the input image, $\mathcal{Q}$ be the set of questions, and q be the input question. a^* represents the predicted answer where $a^* \in \mathcal{A}$, and θ represents the learnable parameters of the model. Then the predicted answer

$$a^* = \underset{a \in \mathcal{A}}{\operatorname{argmax}} P(a|q, i; \theta) \qquad (1)$$

Where $P(a|q, i; \theta)$ represents the probability of an answer given a question and the image. $P(a|q, i; \theta)$ is computed in two stages, namely, the triple filtering stage and the prediction stage.

Triple Filtering Stage: Given a question q and an image i, we retrieve a set of triples $t^* \subset \mathcal{T}$ using an iterative retrieval mechanism, where $\mathcal{T}$ is the whole set of triples in the knowledge graph.

$$t^* = \bigcup_{t \in \mathcal{T}} (t|(P(t|q, i; \theta_r) >= \lambda)) \qquad (2)$$

here, θ_r is the set of learned parameters of the fact retriever module, and λ is a threshold hyper-parameter.

We integrate multi-hop triples, where we specifically focus on utilizing 2-hop triples for contextual information.

Prediction Stage: Then we compute the probability of an answer given question, image and relevant triples as:

$$P(a|i, q) = P(a|t^*, i, q; \theta_p) \quad (3)$$

here, θ_p are the learned parameters of the predictor module and $\theta = \theta_p \cup \theta_r$

3.1 Triple Filtering Module

In this module, we extract relevant information from a large-scale knowledge graph to address questions in KBVQA datasets. It involves two distinct steps:

3.1.1 Triples Relevant to Entities in Image

Our initial step involves extracting image-relevant triples from a vast knowledge graph, effectively reducing dataset size by eliminating unnecessary information. In KVQA and CRIC-VQA, labels representing named entities or object names in the images are available within the dataset, enabling the extraction of relevant triples by identifying all triples with head or tail entities corresponding to these labels. However, in datasets like FVQA lacking inherent labels, we utilize an alternative method detailed in Section 5 to extract relevant triples.

3.1.2 Triples Relevant to Entities in Question

From the refined subset of triples obtained from the first step, the module further refines the triple selection by filtering on the question.

In our approach, inspired by prior work (Wang et al., 2014; Ma et al., 2019; Nayyeri et al., 2023), we leverage embedding similarities to find relevant triples. Preceding triple embedding computation, we substitute all named entities with a <MASK> token. This substitution ensures the model prioritizes predicates over named entities, mitigating the extraction of irrelevant triples. For example, in a query like 'Who is to the right of R.Madhavan?', employing <MASK> prevents irrelevant triples like (R.Madhavan, spouse, Sarita Birje) from being extracted to answer the question.

Number of triples	1	3	5	7	9
Accuracy	68.95%	73.42%	**82.7%**	82.6%	80.20%

Table 1: Exact-match scores when fixed numbers of triples are provided as context for the KVQA dataset.

In contrast to prior studies, which provided a fixed number of similar triples for answer prediction, our approach introduces a dynamic triplet filtering method. We offer the model a variable number of triples based on a similarity threshold criterion. We include triples with similarity scores equal to or greater than the specified threshold. After observation, we found that a threshold of 0.8 effectively captures nearly all relevant triples needed to answer the given questions.

The outcomes of the above approaches are presented in Section 4.

3.2 Prediction Module

To predict the answer based on an image, question, and triples extracted from the triple filtering module, we employ a transformer encoder-decoder model known as OFA (Wang et al., 2022). The complete architecture is depicted in Figure 2, offering a comprehensive overview of our approach. Due to space constraints, we have given the details of the OFA model in Appendix I.

Algorithm 1 outlines the high-level process of retrieving relevant triples and making answer predictions as shown in Appendix E.

4 Experimental Setup & Results

In this section, we explain the results of KVQA and CRIC-VQA datasets.

4.1 Results on KVQA dataset

Table 2 displays the results on the KVQA dataset. Overall, our model exhibits superior performance compared to baseline models and surpasses them in the majority of categories. The KVQA dataset includes 12 classes. However, prior research only made comparisons across 9 classes. Therefore, we also present our results for these 9 classes for a fair comparison. Our model achieves an average score of **85.19%** on the KVQA dataset which is **4.12%** better than the SOTA model POP-VQA. We have also included the results for all 12 classes in the Appendix B. However, we do encounter a major shortfall in the multi-entity class, where our performance is **6%** lower than the current SOTA model, POP-VQA. We attribute this performance gap to the fact that the POP-VQA model is specifically trained for single-hop question-answering contexts. At the same time, our approach incorporates multi-hop triples, potentially introducing additional noise that could affect prediction accuracy.

Previous approaches mainly rely on a fixed number of the most similar triples as context for predictions, our approach employs dynamic filtering, enhancing the model's reasoning capabilities, as shown in Table 3.

Types of Questions	MEMNET	UNITIER	POP-VQA	OFA(Ours) Single-Hop	 Multi-Hop
1-Hop	61.00%	65.70%	**89.80%**	84.25%	86.04%
Boolean	75.10%	94.60%	95.70%	96.89%	**97.17%**
Comparison	50.50%	90.40%	89.60%	**90.82%**	90.15%
Counting	49.50%	79.40%	73.20%	90.08%	**90.32%**
Intersection	72.50%	79.40%	72.30%	87.07%	**89.03%**
Multi-Entity	43.50%	77.10%	**94.90%**	84.01%	88.53%
Multi-Relation	45.20%	75.20%	**93.27%**	90.10%	90.77%
Spatial	48.10%	21.20%	83.89%	92.70%	**94.50%**
Subtraction	**40.50%**	34.40%	37.00%	32.50%	40.20%
Average Scores	53.98%	68.60%	81.07%	83.15%	**85.19%**

Table 2: Results on KVQA (Shah et al., 2019). Exact match scores for various question types. These scores are obtained in a setting where triples are filtered based on both the questions and the images, and the number of triples varies according to a similarity threshold. We show a comparison of our results with the performance of previous baseline models, MEMNET (Tai et al., 2017), UNITIER (Chen et al., 2020) and POP-VQA (Sahu et al., 2024), on the KVQA test set. Bold and underline indicate the best and second-best scores. Overall our model outperforms the baseline across the test set and most of the classes.

4.1.1 Ablation Results

We demonstrate the efficacy of our dynamic filtering method across various settings. Typically, inputs consist of Image Features, Questions, Named Entities, and Context, separated by <SEP> tokens and fed into the transformer encoder. Answers are generated by the transformer decoder. The context is structured as a sequence of triples, labelled as $triple_1$ <SEP> $triple_2$ <SEP> $triple_3$... <SEP> $triple_n$, with triples are in the form (head, relation, tail). These settings include:

1. **No External Knowledge** (Table 3, Row 1): In this setting, we provided image features and questions without any context.
2. **Triples Related to Images** (Table 3, Row 2): Here, we included all triples associated with named entities in the image.
3. **Triple Filtering Based on Questions:** In this context, there exist two configurations, **Fixed Number of Triples** (Table 3, Row 4&6): We choose a fixed number of top-5 triples with the highest similarity scores. While we experimented by varying numbers of triples, as depicted in Table 1, we observed that providing top-5 triples as context yielded the highest accuracy.

   Dynamic Number of Triples with Similarity Threshold (Table 3, Row 3&5): We selected all triples with a similarity greater than or equal to 0.8.

For comparison with baselines on the KVQA dataset, we conducted evaluations on the OFA large model, utilizing a dynamic number of multi-hop triples to determine accuracy across various question classes.

Models	Base	Large
OFA+Image	62.70%	76.70%
OFA+Image+All Triples	72.00%	73.67%
OFA+Image+Filtered Triples(Dynamic) (Single-Hop)	83.65%	85.35%
OFA+Image+Filtered Triples(Top-5) (Single-Hop)	82.45%	83.20%
OFA+Image+Filtered Triples(Dynamic) (Multi-Hop)	**85.15%**	**87.55%**
OFA+Image+Filtered Triples(Top-5) (Multi-Hop)	83.57%	82.70%

Table 3: Ablation Results on the KVQA Dataset. All Triples (Row 2) refers to image-only triple filtering, Filtered Triples involve filtering based on both question and image. In the second approach, two settings are considered: 1) Fixed triples with Top-5 context and 2) Dynamic triples with a similarity threshold. Bold indicates best scores.

4.2 Results on CRIC-VQA dataset

Table 4 presents the results obtained on the CRIC-VQA dataset, which features factual questions requiring commonsense reasoning. Due to the dataset's is not open source, there has been limited prior research. In our study, we compared our method with nine baseline models outlined in (Gao et al., 2023). Our approach achieves an accuracy of **85.80%**, surpassing the SOTA model

Models	Accuracy
Q-Only GRU	55.18%
Q-Only-BERT	59.03%
SAN	63.98%
Bottom-Up+latt	62.39%
MAC-CS	69.65%
NMN-CS	68.96%
Memory-VQA+latt	66.93%
VILBERT+latt	77.54%
VILBERT+ERNIE+latt	79.85%
Ours	
OFA Base (Fixed)	76.17%
OFA Large (Fixed)	79.28%
OFA Base (Dynamic)	81.85%
OFA Large (Dynamic)	**85.80%**

Table 4: Results on CRIC-VQA (Gao et al., 2021). Exact match scores for various baselines as well as our model. Fixed denotes fixed number of triples with Top-5 context, and dynamic denotes variable triples with a similarity threshold.

VILBERT+ERNIE+latt by **5.95%**. As the number of baselines is high, we explain each in the Appendix G.

The CRIC-VQA dataset possesses a relatively small knowledge base, containing approximately **3,400** triples. To demonstrate the effectiveness of our method, we expanded the knowledge base using ConceptNet. Our augmentation involved incorporating all triples related to the objects depicted in the images, ensuring alignment with either the head or tail entity corresponding to the object label. Consequently, we augmented the knowledge base to a substantial **99,586** triples. This significant increase presents challenges in extracting relevant knowledge for question-answering tasks. Given the dataset's size and computational constraints, our experiments primarily focused on filtering context based on both images and questions.

Due to space constraints, training details are present in Appendix A.

5 Generalisation Capability

We demonstrate our model's generalization capability by fine-tuning it on the FVQA dataset following pretraining on the KVQA dataset. The primary challenge with the FVQA dataset is the absence of object labels within the dataset itself. So extracting image-relevant triples directly from the knowledge graph by matching the head or tail entity is not possible. We fine-tuned the CLIP model (Radford et al., 2021) to get image-relevant triples. We have included details of fine-tuning the CLIP model in Appendix D. To find image-relevant triples we calculate the CLIP embedding (Radford et al., 2021) for each triple. To ensure we extract relevant triples for small objects in the image, we divide the image into four equal-sized patches and compute the CLIP embedding for each patch. When examining the entire image without dividing it into patches, important details related to small objects (such as the flower vase) as shown in Figure 3 might be overlooked. Cosine similarity between patch embeddings and all triples is calculated, selecting those with a similarity above 0.8 for each patch. For extracting the triples relevant to the question we use the same approach as explained in Section 3.1.2. Our approach achieves an accuracy of **65.28%**, surpassing the SOTA by **4.28%**. In Table 5, we compare our work with previous baselines, particularly those that do not consider named entities when extracting relevant triples from the knowledge base. This improvement is attributed to our model's ability to eliminate irrelevant noise introduced by external context, unlike the SOTA model MM-Reasoner, which integrates context from diverse sources such as image captions, GPT4 and many more.

Models	Accuracy
Human	77.99%
FVQA (Wang et al., 2018)	56.91%
ZS-FVQA (Chen et al., 2021)	58.27%
FVQA (Ensemble) (Wang et al., 2018)	58.76%
MM-Reasoner (Ensemble) (Khademi et al., 2023)	61.10%
Ours	
OFA Base(Ours)	54.00%
OFA Large(Ours)	**65.28%**

Table 5: Results on FVQA. Exact match scores for various baselines as well as our model. Utilized pre-trained model on KVQA dataset under dynamic multi-hop setting (Table 3). The inference is done while providing the dynamic number of triples as context.

Ablation Results : We explore various settings to demonstrate the effectiveness of incorporating a dynamic number of triples during pretraining or fine-tuning, as depicted in Table 6. The table showcases results with and without fine-tuning the FVQA dataset, presenting different contextual information during these processes. Notably, we find that utilizing a dynamic number of triples leads to a **12%** performance improvement compared to a fixed number. We also include results without image segmentation into patches and results ob-

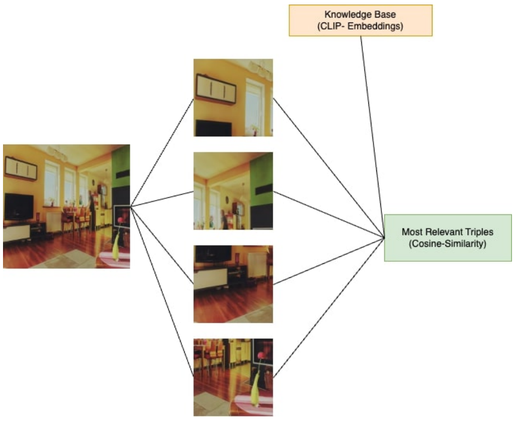


Figure 3: Splitting the image into four patches to extract relevant triples.

tained by identifying triples relevant to objects in the image. For this, we utilize the object detection model (Wu et al., 2019) to obtain bounding boxes for objects and extract the most relevant triples for each box. For a comprehensive understanding of these approaches, including detailed results, please refer to Appendix C. Despite the distinct domains between the fine-tuning dataset (FVQA) and the pretraining dataset (KVQA), our model exhibits strong performance and generalizability across different domains.

	Context-Type			
Models	**Pre-training**	**Inference**	**∼ FT**	**FT**
Base	fixed	fixed	20.35%	43.00%
Base	fixed	dynamic	21.90%	30.00%
Large	fixed	fixed	36.48%	39.00%
Large	fixed	dynamic	38.70%	50.00%
Base	dynamic	fixed	34.50%	41.51%
Base	dynamic	dynamic	40.14%	54.00%
Large	dynamic	fixed	43.50%	58.00%
Large	dynamic	dynamic	**47.00%**	**65.28%**

Table 6: Ablation results for FVQA dataset: Exact match scores comparing fine-tuned (FT) and non-fine-tuned (∼FT) models, pre-trained on the KVQA dataset. The pre-training context type specifies how the model was trained on the KVQA dataset, while the inference context type indicates the settings for fine-tuning and inference on the FVQA dataset.

6 Relevance of Knowledge in the Context of MLLMs

Given the extensive training of MLLMs on vast datasets, it's natural to assume that external knowledge might not be essential for using them in tasks like KB-VQA. Recently, there have been extensive discussions about whether Multimodal Large Language Models (MLLMs), trained on large datasets, can answer KB-VQA questions based solely on their internal knowledge or if external information is necessary. In this section, we illustrate that relying solely on the implicit knowledge within MLLMs is insufficient for addressing such questions. Additionally, we'll demonstrate the effectiveness of our knowledge retrieval method by evaluating its performance with a contemporary vision language model.

6.1 Zero-shot Evaluation on the LLAVA model

Dataset	**Without External Knowledge**	**With External Knowledge**
KVQA	55.20%	64.50%
CRIC-VQA	58.60%	69.40%

Table 7: Zero-shot results on LLAVA model: Exact Match scores achieved by the llava-v1.6-vicuna-13b model. The results are reported for two settings: (1) without providing any prior knowledge and (2) with the inclusion of knowledge in the form of triples, alongside questions and named entities found in the images.

We conducted experiments with the MLLM llava-v1.6-vicuna-13b[1], prompting it to generate responses to zero-shot image prompts under two conditions: one without external knowledge and the other with external knowledge obtained through our dynamic triple retrieval module. The prompts for both conditions are detailed in Appendix F. For evaluation, we computed exact match scores by comparing generated answers with correct ones. To determine whether discrepancies arose from the model's output or it is exact match metric issues, we performed a qualitative analysis on 200 incorrect samples. Additional information regarding this is provided in Appendix H.

The results of the above approach are shown in Table 7. We can observe incorporating explicit external knowledge increased accuracy by approximately **10.05%**. The significant increase in accuracy demonstrates how even large language models benefit from incorporating external knowledge.

6.2 Finetuning LLAVA on KVQA dataset

Here we demonstrate the effectiveness of our approach while finetuning LLAVA for the KB-VQA task. We fine-tuned LLAVA in three different settings:

[1] https://huggingface.co/liuhaotian/llava-v1.6-vicuna-13b

1. **Without Knowledge**: Here no context is provided to answer the question.
2. **With Fixed Knowledge**: Here top-5 most similar triples are provided as context to answer the question.
3. **With Dynamic Knowledge**: Here triples obtained from the dynamic triple retrieval module are provided to answer the question.

The results are shown in Table 8. There is a substantial increase in accuracy **15.7%**, on the KVQA test set when contextual information is integrated while fine tuning. This accuracy gap further widens to **20.2%** when the triples are filtered using our dynamic filtering approach. These results underscore the effectiveness of our approach, even with recent models, significantly enhancing prediction accuracy.

Technique	**Accuracy**
LLAVA+Labels	72.40%
LLAVA+Labels+Knowledge (fixed)	88.10%
LLAVA+Labels+Knowledge (dynamic)	**92.60%**

Table 8: Results of LLAVA on KVQA Dataset: Exact Match scores achieved by the LLAVA model after fine-tuning on the KVQA Dataset for the KB-VQA task. In this context, **labels** refer to named entities detected within the images, while **knowledge** indicates external information supplied either statically or dynamically.

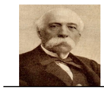
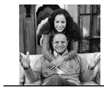
Image	**Question**	**Truth Value**	**No Triples**	**All Triples**	**Filtered Triples**
	Is the person in the image a politician?	No	Yes	No	No
	For how many years did the person in the image live?	83	72	82	83
	Were all the people in the image born in the same country?	No	Yes	Yes	No
	Who among the people in the image ever married Vladimir Soshalsky?	Person on the left	Person on the right	Person on the right	Person on the right

Table 9: Qualitative analysis, which presents instances from the dataset and their answer predictions with and without the presence of triples in the input.

7 Qualitative Analysis

We'll illustrate how integrating knowledge boosts the OFA model's predictive power while efficiently filtering noise enhances its reasoning capability. We choose some samples as shown in Table 9. The first row shows that giving extra information helps the model make correct predictions for simple questions with straightforward answers. In this case, the correct answer is obtained whether or not we filter the knowledge triples. Rows 2 and 3 demonstrate that providing all triples filtered based on image and not question introduces irrelevant knowledge (noise), resulting in inaccurate predictions. However, filtering triples based on questions leads to correct answers. These complex questions include a single-hop subtraction (Row 2) and a multi-hop boolean query (Row 3), requiring noise removal for accurate predictions.

In the final example, regardless of whether knowledge triples are supplied or not, the model produces incorrect answers. This question falls under the spatial and multi-hop categories, requiring the model to make inferences based on both image features and external knowledge. One possible explanation for the inaccurate predictions could be the model's insufficient training to a broad range of questions within these complex categories, hindering its ability to reason effectively in such scenarios. Due to space constraints, we provide these and some more examples in Table 16.

8 Conclusion

We presented a novel approach for KBVQA, utilizing a dynamic triple-filtering module to extract external context from knowledge graphs. Our method outperforms the SOTA on three different KBVQA datasets, achieving an average improvement of **4.75%**. We demonstrated that providing the model with a varying number of triples during pre-training or fine-tuning enhances its reasoning capabilities compared to a fixed number of triples. We also showcased the generalization capability of our approach by achieving SOTA performance on a small dataset using a model trained on a completely different domain. Furthermore, we demonstrated that large MLLMs also require external knowledge for accurate responses. Finally, we also demonstrated the effectiveness of our approach on the latest MLLM LLAVA, emphasizing that providing a dynamic number of triples improves accuracy by **20%** as compared to static ones. The key insight is that dynamically determining the number of relevant triples in the context eliminates noise, resulting in more precise predictions. We also discuss some future explorations in Appendix K.

References

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning.

Zhuo Chen, Jiaoyan Chen, Yuxia Geng, Jeff Z. Pan, Zonggang Yuan, and Huajun Chen. 2021. Zero-shot visual question answering using knowledge graph.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

D. Gao, R. Wang, S. Shan, and X. Chen. 2023. Cric: A vqa dataset for compositional reasoning on vision and commonsense. *IEEE Transactions on Pattern Analysis &amp; Machine Intelligence*, 45(05):5561–5578.

Difei Gao, Ruiping Wang, Shiguang Shan, and Xilin Chen. 2021. Cric: A vqa dataset for compositional reasoning on vision and commonsense.

Diego Garcia-Olano, Yasumasa Onoe, and Joydeep Ghosh. 2021. Improving and diagnosing knowledge-based visual question answering via entity enhanced knowledge injection.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. Kat: A knowledge augmented transformer for vision-and-language.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging.

Mahmoud Khademi, Ziyi Yang, Felipe Frujeri, and Chenguang Zhu. 2023. MM-reasoner: A multi-modal knowledge-aware framework for knowledge-based visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6571–6581, Singapore. Association for Computational Linguistics.

Paul Lerner, Olivier Ferret, and Camille Guinaudeau. 2024. Cross-modal Retrieval for Knowledge-based Visual Question Answering. Working paper or preprint.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, Jose G. Moreno, and Jesús Lovón Melgarejo. 2022. Viquae, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3108–3120, New York, NY, USA. Association for Computing Machinery.

Guohao Li, Xin Wang, and Wenwu Zhu. 2020. Boosting visual question answering with context-aware knowledge aggregation. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1227–1235, New York, NY, USA. Association for Computing Machinery.

Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. 2022. Revive: Regional visual representation matters in knowledge-based visual question answering.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Minbo Ma, Fei Teng, Wen Zhong, and Zheng MA. 2019. A sentence-rcnn embedding model for knowledge graph completion. In *2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, pages 484–490.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge.

Mojtaba Nayyeri, Zihao Wang, Mst. Mahfuja Akter, Mirza Mohtashim Alam, Md Rashad Al Hasan Rony, Jens Lehmann, and Steffen Staab. 2023. Integrating knowledge graph embeddings and pre-trained language models in hypercomplex spaces. In *22nd International Semantic Web Conference (06/11/23 - 10/11/23)*.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, et al. 2023. Gpt-4 technical report.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Pragya Paramita Sahu, Abhishek Raut, Jagdish Singh Samant, Mahesh Gorijala, Vignesh Lakshminarayanan, and Pinaki Bhaskar. 2024. Pop-vqa - privacy preserving, on-device, personalized visual question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8470–8479.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8876–8884.

Violetta Shevchenko, Damien Teney, Anthony Dick, and Anton van den Hengel. 2021. Reasoning over vision and language: Exploring the benefits of supplemental knowledge.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2018. Conceptnet 5.5: An open multilingual graph of general knowledge.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations.

Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. 2017. Memnet: A persistent memory network for image restoration.

Peter Vickers, Nikolaos Aletras, Emilio Monti, and Loïc Barrault. 2021. In factuality: Efficient integration of relevant facts for visual 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 2: Short Papers)*, pages 468–475, Online. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2018. Fvqa: Fact-based visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(10):2413–2427.

Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2017. Fvqa: Fact-based visual question answering.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph and text jointly embedding. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1591–1601, Doha, Qatar. Association for Computational Linguistics.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

A Training Details

The experiments encompassed both OFA Base and Large models, maintaining image resolutions at 480×480 and 640×640 for Base and Large models, respectively. The dropout rate was set at 0.1. Adam Optimizer was employed with beta values of 0.9 and 0.999, epsilon set to 1×10^{-08}, and a warm-up ratio of 0.06. An initial learning rate of 1×10^{-5} with polynomial decay was utilized. During test inference, a beam size of 10 and a temperature of 0.98 were applied. T5-Base (Raffel et al., 2023) model generated embeddings for questions and triples in the triple filtering process. The training was conducted on Nvidia RTX A6000 [2], with each iteration taking approximately 8 and 12 hours for the Base and Large models, respectively. We utilized the LLAVA model for fine-tuning the KVQA dataset for the KB-VQA task. Fine-tuning was conducted with a learning rate of 2×10^{-5} over 2 epochs, with a warmup ratio set to 0.03. All experiments were performed on three Nvidia RTX A6000 GPUs, with each iteration requiring approximately 18 hours to obtain conclusive results.
The number of parameters used and the number of encoder-decoder layers for both the OFA-Base and OFA-Large models are given in Table 11.

B Additional Results for KVQA Dataset

In Section 4, we demonstrated the results for 9 classes on the KVQA dataset, aligning with the prior state-of-the-art model, POP-VQA (Sahu et al., 2024). However, in this Section, we extend our analysis to cover all 13 classes within the KVQA dataset, as detailed in Table 10. Additionally, we include the results obtained from the MEMNET (Tai et al., 2017) and UNITIER (Chen et al., 2020) models for a fair comparison.

We also present results for the other two scenarios. First, no triples are given as context, second when we include all the triples associated with the image, without any filtering based on the question as explained in Section A. The results for these approaches can be found in Table 12.

In our observations, it becomes evident that including all triples results increase in accuracy across most categories when compared to not including any triples at all. However, in more complex categories such as subtraction, the accuracy improvement is not as significant, mainly because accurate predictions demand more refined triples.

An interesting observation occurs when we look at the spatial category. When we provide all triples, accuracy decreases, indicating that in the spatial category, the inclusion of triples is unnecessary. This result shows that our dynamic triple extraction module works effectively, especially in spatial questions, where it rarely provides external triples. This emphasizes that the module can smartly adjust to meet the specific needs of each question.
We discuss this in Section 4.1

C Additional Results for FVQA Dataset

In Section 5, due to the absence of labels in the dataset, we utilized the CLIP model to extract relevant triples from the image. To achieve this, we divided the image into four patches, computing the most relevant triples for each patch. In this Section, we present results for two additional settings to ensure transparency,

Triples relevant to Full Image: In this configuration, we refrain from dividing the image into patches. Instead, we compute relevant triples for the entire image. These results are summarized in Table 13. The problem with this approach is that when computing the cosine similarity of the CLIP embedding of the entire image and triples, triples relevant to smaller objects might not be captured. For instance, as depicted in Fig 3, triples related to the flower vase could be overlooked.

Triples relevant to objects in the image: For extracting the triples relevant to objects in the image we use the following approach:

- Bounding Box Extraction: We identify bounding boxes for each object present in the image. These bounding boxes define the spatial regions corresponding to the objects.

- Detectron Model: To achieve this, we utilize the Detectron model, which detects the precise coordinates of the bounding boxes.

- Image Patch Extraction: Once we have the bounding box coordinates, we extract image patches corresponding to those regions.

- Triple Extraction: For each image patch, we find the relevant triples associated with the objects within that patch.

The results are demonstrated in table 14.

[2] https://www.nvidia.com/en-in/design-visualization/rtx-a6000/

Types of Questions	MEMNET	UNITIER	OFA(Ours) Single-Hop	 Multi-Hop
1-Hop	61.00%	65.70%	84.25%	86.04%
1-Hop Counting	-	78.0%	88.80%	**90.74%**
1-Hop Subtraction	-	28.60%	31.25%	**37.89%**
Multi-Hop	53.20%	87.90%	60.80%	**90.40%**
Boolean	75.10%	94.60%	96.89%	**97.17%**
Comparison	50.50%	90.40%	**90.82%**	90.15%
Counting	49.50%	79.40%	90.08%	**90.32%**
Intersection	72.50%	79.40%	87.07%	**89.03%**
Multi-Entity	43.50%	77.10%	84.01%	88.53%
Multi-Relation	45.20%	75.20%	90.10%	90.77%
Spatial	48.10%	21.20%	92.70%	**94.50%**
Subtraction	**40.50%**	34.40%	32.50%	40.20%

Table 10: The table displays the results of all 13 classes on the KVQA dataset. These scores are obtained in a setting where triples are filtered based on both the questions and the images, and the number of triples varies according to a similarity threshold.

Model	#Param	#Enc.Layers	#Dec.Layers
OFA-Base	182M	6	6
OFA-Large	472M	12	12

Table 11: The table displays information regarding the parameter count, as well as the number of encoder and decoder layers for both the OFA Base and OFA Large models.

Types of Questions	OFA(Ours) With No Triples	 With All Triples
1-Hop	72.20%	76.81%
1-Hop Counting	75.95%	76.00%
1-Hop Subtraction	29.80%	30.06%
Boolean	86.10%	94.40%
Comparison	83.59%	88.77%
Counting	81.10%	81.30%
Intersection	78.19%	76.40%
Multi-Entity	71.10%	76.32%
Multi-Hop	74.22%	81.70%
Multi-Relation	72.12%	83.92%
Spatial	89.02%	83.43%
Subtraction	4.50%	7.20%

Table 12: The table presents the performance of various question types in two distinct scenarios: one without the inclusion of any triples as context (referred to as "With No Triples"), and the other with all the relevant triples filtered by images, while not applying any filtering on the questions (referred to as "With All Triples").

In the above two approaches, we filtered the triples based on the image, for further filtering based on a question we used the same method as explained in Section 3.1.2.

For prediction we employed a pre-trained model on the KVQA dataset, specifically focusing on the best setting where the model was trained with multi-hop dynamic triples as context. The fine-tuning and inference process also considers a dynamic number of triples as context. We have provided results for both scenarios: without and with fine-tuning on the FVQA dataset, as elaborated in Section 5.

D Training CLIP Model

CLIP model (Radford et al., 2021) is trained for image-text similarity and not for image-triples similarity. Therefore, we train the CLIP model to extract triples that are relevant to the image. We denote the set of triples from the knowledge graph as t_k, and the reference image as *I*. To identify the triples that are relevant to the reference image, we minimise the following objective,

$$-\log \frac{\exp(s(I, t_k^{(+)})e^{\tau})}{\exp(s(I, t_k^{(+)})e^{\tau}) + \sum_j \exp(s(I, t_k^{(j)})e^{\tau})}$$

We implement $s(I, t_k^{(+)})$ using CLIP as:
$s(I, t_k^{(+)}) = \cos(CLIP_V(I), CLIP_T(t_k))$
Here $t_k^{(+)}$ denotes the triple relevant to the image, $t_k^{(j)}$ denotes the irrelevant triples for an image and τ denotes temperature parameter which

Model	Without-fine-tuning	With-fine-tuning
OFA-Base	33.28	39.94
OFA-Large	34.84	43.20

Table 13: Results on FVQA dataset. Exact match score with and without fine-tuning on the FVQA dataset. Triples relevant to images are computed by considering the whole image without dividing it into patches.

Model	Without-fine-tuning	With-fine-tuning
OFA-Base	33.75	44.62
OFA-Large	38.42	46.71

Table 14: Results on FVQA dataset. Exact match score with and without fine-tuning on the FVQA dataset. Triples relevant to images are computed by considering each object in the image.

controls the range of the logits in the softmax as explained in (Radford et al., 2021). Since there isn't a specific dataset available for images and their relevant triples, we utilize the ViQuae Wikipedia Corpus (Lerner et al., 2022) to acquire the images and their corresponding triples. We have chosen 2000 instances that include images and their related triples, which were extracted using the Wikidata knowledge graph (Vrandečić and Krötzsch, 2014). We train the CLIP model using the above objective to get relevant triples.
We discuss this in Section 5

E Algorithm

The algorithm is discussed in 1

F Prompting on LLAVA Model

In this segment, we'll furnish the prompt utilized to find responses from the llava-hf/llava-v1.6-mistral-7b-hf model for image-related questions. To ensure fair comparison based on exact match scores, we want concise answers to avoid any extraneous information. The provided prompt generates concise responses, minimizing any potential noise.

Prompt for evaluation without giving any knowledge
Please answer concisely in one or two words:
Question: <question>
Named Entities: <named entities>

Prompt for evaluation when giving knowledge
Please answer the question concisely in one or two words. We also provide Named Entities and knowledge triples separated by <sep> token for your assistance:
Question: <question>
Named Entities: <named entities>
Triples: <triples string>
We discuss this in Section 6

Algorithm 1 Retrieving context for k-hop Question Answering and feeding the Question, Image, and Context into a Transformer Encoder-Decoder model to predict the desired answer.

```
Require:
  $Q_0 \rightarrow$ Input Question
  $T \rightarrow$ Triples from Knowledge Graph
  $k \rightarrow$ Number of Hops
  $I \rightarrow$ Image
  $E \rightarrow$ Named Entities
Ensure:
  Triple Filtering (By Images)
  for $Count$ in $k$ do
      for $(Head, Relation, Tail)$ in Knowledge Graph do
          if Head or Tail in $E$ then
              Relevant Triples += $(Head, Relation, Tail)$
          end if
      end for
  end for
  Triple Filtering
  for Triple in Relevant Triples do
      $T\_Embed$ = T5 Base(Triple)
      $Q\_Embed$ = T5 Base($Q_0$)
      if Similarity(T_Embed, Q_Embed) $\geq \lambda$ then
          Context += Triple
      end if
  end for
  Prediction Module
  Answer = OFA_Model($Image < SEP > Question < SEP > NamedEntities < SEP > Context$)
```

G Baselines on CRIC-VQA dataset

In this section, we explain each baseline in brief as depicted in Table 4.

Q-Only GRU - Q-Only model only takes the GRU question features as input.

Q-Only BERT - Q-Only model only takes the BERT question features as input.

SF - SF first uses visual concepts extracted by object, scene, action predictors, CNN image feature, and LSTM question feature to retrieve the Top-1 related knowledge item, then uses the question feature and retrieved knowledge item to predict the answer.

Bottom-Up+latt - Bottom-Up is a traditional VQA model emphasizing object-level reasoning with soft attention to object regions. This baseline enhances Bottom-Up by incorporating a binary cross-entropy loss on attention scores to guide the model to focus on the correct region when combining attended image and question features for generating the final answer.

MAC-CS - MAC is a leading modular VQA model designed for CLEVR and GQA. It breaks down questions into attention-based reasoning steps. The expanded MAC's capabilities to incorporate access to knowledge items resulted in MAC-CS, which focuses on commonsense reasoning.

NMN-CS - The Neural Modular Network (NMN) is a distinct VQA model. However, its original iterations are not directly applicable to commonsense questions. To address this limitation, visual commonsense reasoning modules have been integrated, resulting in NMN-CS.

Memory-VQA+latt - This memory network operates by encoding input materials such as knowledge items and the image in the CRIC as memories. It utilizes the question to initiate an iterative attention process, enabling the model to retrieve relevant information for answering the question. In contrast to Memory-VQA, this baseline further incorporates a cross-entropy loss on attention scores.

VILBERT+ERNIE+latt - The model consists of three modules: ViLBERT for image and question feature extraction, ERNIE for candidate knowledge item feature extraction, and an attention module for predicting the answer by using pooled features from both transformers to locate the target image region.

We discuss this in Section 4.2

H Qualitative Analysis of LLAVA Generated Answers

We analyzed LLAVA-generated answers on the KVQA dataset in a zero-shot scenario. The primary objective was to confirm whether the model's incorrect answers were a result of its output or an issue with the exact match metric. To achieve this, we sampled a total of random 200 instances containing questions of all 13 classes where the LLAVA model provided incorrect answers. We determined whether the incorrect answers were due to the model itself or a metric-related problem. This evaluation yielded counts for both scenarios: instances where the model's answers were incorrect and instances where the issue lay with the metric. The results of the evaluation are shown in Table 15. The primary issue arises with spatial inquiries where the correct response is "Person on the Left", or "Person on the Right", or "Person on the Center" yet the model tends to provide named entities of individuals instead.

Model mispredictions	Metric problem
182	18

Table 15: Human Evaluation results for the LLAVA Model Output

For example:
Question: Who among the people in the image lived longest?
Truth Answer: Person in the left
Predicted Answer: Lili Damita

For such questions, given the limited options of only three potential answers, we adjust the prompt as:

Updated prompt for Spatial Class:
Please answer concisely in one or two words:
Question: <question>
Named Entities: <named entities>
Don't give named entities in the answer instead provide the answer in form Person in Center, Person in Left, Person in Right.

The outcomes presented in Table 7 account for this scenario to guarantee a fair assessment process.

Regarding the CRIC-VQA dataset, which focuses on objects and typically elicits responses of one or two words like "Desk", "Water" etc there is no issue with metric and the model generates wrong answers.

We discuss this in Section 4.2

I OFA Model

We leverage the power of Unified Vision-Language (VL) modelling (Wang et al., 2022), which has demonstrated significant potential across various VL tasks. For our VQA tasks, we adopt a vision language transformer encoder-decoder model OFA Base and OFA Large architecture. The OFA model is designed to handle diverse tasks and modalities, seamlessly integrating vision-only, language-only, and vision-language tasks within a sequence-to-sequence learning framework.

Our input comprises ResNet152 (He et al., 2015) features extracted from the image, followed by the question and context, both tokenized using byte-pair encoding (BPE) (Bostrom and Durrett, 2020). We employ a unified vocabulary that encompasses tokens from both visual and linguistic domains. Transformers serve as the core encoders and decoders, treating the vision-language task as a sequence-to-sequence problem.

J More Examples

Refer to Table 16 for the examples used in Section 7, as well as some additional examples that demonstrate the effectiveness of our approach. Table 16 includes certain questions that do not necessitate any knowledge (as seen in Row 7). These can be addressed solely based on image features, without the need for external knowledge. Supplying triples in these instances results in incorrect predictions. These questions predominantly belong to the spatial category. Additionally, some questions are straightforward and do not require knowledge filtering (as seen in Row 10). Providing all triples without filtering based on questions in these cases would also yield correct answers, eliminating the need for filtering. These questions are primarily 1-hop questions. However, for complex categories such as 1-hop subtraction, multi-hop, etc., a robust reasoning capability is required. Therefore, supplying filtered knowledge is essential to prevent any confusion that could lead to incorrect predictions. We discuss this in Section 7

K Future Work

Several potential avenues for future exploration are available. Presently, the fact retriever and answer prediction module undergo separate training processes. Exploring an end-to-end trainable model that seamlessly integrates both components represents an intriguing direction to explore. The optimal number of triplets for context was determined through experimentation, incorporating heuristics for similarity values, among other factors. However, enhancing performance can be achieved through the model's automatic learning of this ideal number of triplets based on the characteristics of the question, image, etc. Exploring additional techniques to enhance the model's generalization across different domains is another compelling direction to investigate. Creating an explanatory model for the retrieved context would prove beneficial for numerous practical applications. We anticipate that the numerous avenues for future work, along with our presented results, will inspire further exploration and advancements in the KBVQA domain.

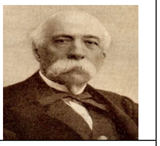
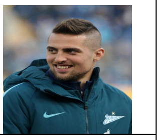
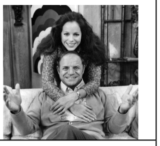
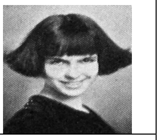
Question	True Answer	No Triples	All Triples	Filtered Triples	Image
Is the person in the image a politician?	No	Yes	No	No	
In which country was the person in the image born?	Slovakia	Hungary	Slovakia	Slovakia	
For how many years did the person in the image live?	83	72	82	83	
Were all the people in the image born in the same country?	No	Yes	Yes	No	
Who among the people in the image ever married Vladimir Soshalsky?	Person on the left	Person on the right	Person on the right	Person on the right	
For how many years did the person in the image live?	79	86	85	79	
Do all the people in the image have a common occupation?	No	Yes	Yes	No	
Who is to the right of Jorge Toriello Garrido?	Jacobo Árbenz	Jacobo Árbenz	jajaxedlol	Jacobo Árbenz	
In which year did the person in the image start professional activities?	1911	1920	1986	1956	
Who among the people in the image ever married to Bill Williams?	Person in the right	Person in the left	Person in the right	Person in the right	

Table 16: Error analysis table, presents instances from the datasets and their predicted answers in three settings mainly no triples, all triples and filtered triples.