Precision in Knowledge Empowers, Excess is Distraction: Visual Question Answering with Knowledge-Infused Language Models

Anonymous ACL submission

Abstract

In the realm of multimodal tasks, Visual Question Answering (VQA) plays a crucial role by addressing questions in natural language grounded in visual content. Knowledge-Based Visual Question Answering (**KBVQA**) elevates this concept by integrating external knowledge with images to respond to questions. KBVQA shows great potential in tackling real-world challenges, encompassing assistance for the visually impaired and enhancing image search functionalities. We introduce an innovative 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 pertinent 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 SOTA on **three** different KBVQA datasets. Through thorough experiments and analysis, we illustrate that furnishing variable triples for each question *improves the reasoning capabilities of the language model* in contrast to supplying a fixed number of triples. 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. A typical KBVQA model consists of both a knowledge retrieval module and a predictor module. 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, without requiring external information

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, furnishing a consistent number of triples for all questions may lead to either inadequate informa-

tion or unnecessary noise, potentially resulting in inaccurate predictions. E.g. in Figure 1, *Who is to the right of Ion Antonescu?* 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. 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,

- An innovative approach for KBVQA, featuring a **dynamic triple filtering** method, provides varying question-specific triples instead of a fixed number, serving as context to address the posed question.
- The approach surpasses the state-of-the-art (SOTA) on three different KBVQA datasets. On the KVQA dataset (Shah et al., 2019), it achieves 85.19% accuracy, marking a **4.12%** improvement over the previous SOTA. On the CRIC-VQA dataset (Gao et al., 2023), it reaches 85.80% accuracy, surpassing the previous SOTA by **5.95%**. On the FVQA dataset (Wang et al., 2017), the method attains 65.28% accuracy, exceeding the prior SOTA (Khademi et al., 2023) by **4.18%**
- The outcomes set a fresh 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."
- We also present an improved 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. This expansion serves to further challenge the knowledge extraction capabilities of upcoming models.

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, exemplified 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.

2.1 Datasets

We have used three main datasets for Closed Domain Knowledge-Based VQA. KVQA (Shah et al., 2019), CRIC-VQA (Gao et al., 2023) and FVQA (Wang et al., 2017). The KVQA dataset is an extensive set with 183,000 question-answer pairs, over 18,000 distinct named entities, and a collection of 24,000 images. The questions in this dataset require an understanding of the named entities present in the image. Conversely, the FVQA and CRIC-VQA datasets lean more towards commonsense understanding rather than focusing on named entities in the image. The FVQA dataset contains 5826 questions and 2190 associated images and CRIC-VQA comprises the larger dataset with 494K questions and 94K images. All datasets require external knowledge to answer the question. We selected these three datasets as they cover a broad spectrum of questions, including those based on user-centric questions and general factual questions. For getting external knowledge, we mainly utilized the Wikidata (Vrandečić and Krötzsch, 2014) and ConceptNet (Speer et al., 2018) knowledge graphs.

3 Our Approach

Our approach follows a two-stage process to determine the answer to 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

$$a^* = argmax_{a \in \mathcal{A}} P(a|q, i; \theta) \quad (1)$$

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.

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

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 and image 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

This module serves the purpose of extracting 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 first step involves extracting triples relevant to images from an extensive knowledge graph. This significantly reduces the dataset size, removing superfluous information. In the KVQA and CRIC-VQA datasets, labels 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, where labels are not inherently available, we adopt an alternative approach to extract triples relevant to images, as elaborated in Section 5.

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.

As in previous works (Wang et al., 2014; Ma et al., 2019; Nayyeri et al., 2023) we use embedding similarity to find the relevant triples. For this, we employ the T5-Base model (Raffel et al., 2023) to generate embeddings for both the question and all the triples. Before calculating the triple embeddings, we substitute all named entities in the triples with a <*MASK*> token. Substituting named entities with the <*MASK*> token ensures that the model focuses on predicates rather than focussing on named entities. This prevents the extraction of irrelevant triples. For instance, consider the question: *Who is to the right of Ion Antonescu?* In this scenario, the model does not need any triple to answer the question. It can be answered from the image features

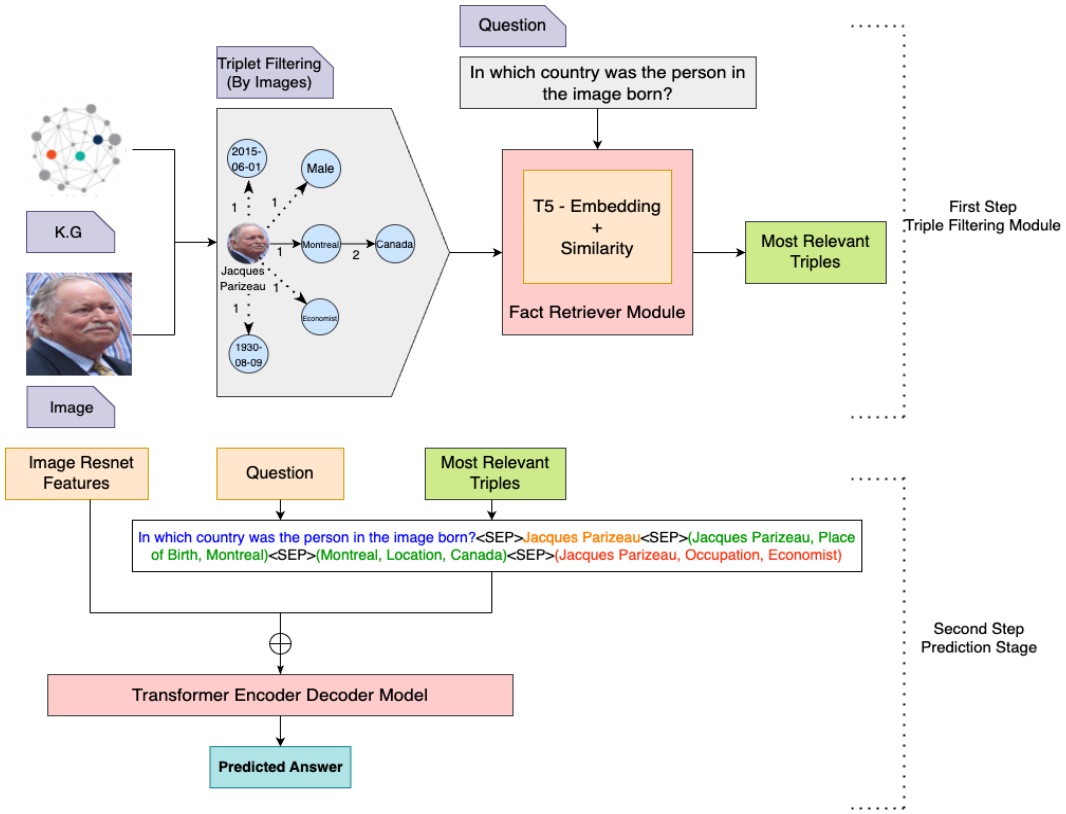


Figure 2: The proposed framework is illustrated in the flow diagram. In the first stage of prediction, a subset of triples is extracted from the image and knowledge graph, followed by an additional round of filtering where these triples are filtered based on the similarity between the question embedding and triple embedding obtained using the T5 model. In the second stage of prediction, **Relevant Triples**, **Image Resnet Features** and **Question**, are fed into a transformer encoder-decoder model (OFA) to generate the predicted answer.

itself. But when <*MASK*> token is not employed, the model might extract several irrelevant noisy triples associated with Ion Antonescu, for example, (Ion Antonescu, sex, male).

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

Table 1: The table displays the accuracy corresponding to varying numbers of triples provided as context for the KVQA dataset.

We conducted experiments in two different settings based on the number of triples provided as context to the predictor module.

Fixed Number of Triples: In this scenario, we opt for a predefined quantity of the most comparable triples to function as context for predicting answers.

Dynamic Number of Triples: In this methodology, we furnish the model with a flexible number of triples as contextual information. We incorporate triples with a similarity score greater than or equal to the specified threshold. Through observation, we determined that a threshold value of 0.8 proves to be the most effective, encompassing nearly all the triples necessary to address the given question.

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.

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 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.

Algorithm 1 outlines the high-level process of retrieving pertinent triples and making answer predictions.

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($Image < SEP > Question < SEP > NamedEntities < SEP > Context$)

4 Experimental Setup & Results

We conducted experiments in different settings, as outlined in Table 4.

Without Infusion & With Infusion: In the without infusion setting, we do not provide any external triples to the OFA model. The input format consists of Image Features, Questions, and Named Entities separated by <SEP> directly to the transformer encoder. The corresponding answers are then generated by the transformer decoder model. The OFA model's vision-language understanding demonstrates the capability to accurately predict certain answers. In the Infusion setting, we include context along with the image and question. Here, context refers to all the relevant triples from the knowledge graph that help in answering the question. The input format consists of Image Features, Questions, Named Entities, and Context, separated by <SEP>. This format presents context as a sequence of triples, labeled as $triple_1$ <SEP> $triple_2$ <SEP> $triple_3$... <SEP> $triple_n$.

The results for KVQA dataset are shown in Table 4. We employed diverse settings to illustrate that providing a dynamic number of triples based on a similarity threshold, as opposed to a fixed number, significantly enhances accuracy. We explored various experimental settings as follows,

1. **No External Knowledge** (Table 4, Row 1): In this setting, we provided image features and questions without any context.
2. **Triples Related to Images** (Table 4, 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 4, 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. Consequently, we opted to utilize the $Top-5$ triples as context for our analysis.
   Dynamic Number of Triples with Similarity Threshold (Table 4, 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 as shown in Table 2. 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 accuracy of 85.19% on the KVQA dataset which is **4.12%** better than SOTA model POP-VQA (Sahu et al., 2024). We have included the results for all 12 classes in the Appendix B.

We also apply our approach to the CRIC-VQA dataset, which comprises factual questions requiring commonsense reasoning. The knowledge base

Types of Questions	MEMNET	UNITIER	POP-VQA	OFA(Ours) Single-Hop	OFA(Ours) 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.

Models	Accuracy
Q-Only GRU	55.18%
Q-Only-BERT	59.03%
SF	60.98%
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 3: **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.

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 4: **Various Setting Results on KVQA**. All Triples (Row 2) refers to image-only triple filtering, Filtered Triples involves 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.

for CRIC-VQA is relatively small, with around 3,400 triples. To showcase the efficacy of our approach, we enhance the knowledge base using ConceptNet. We incorporate all triples associated with the objects depicted in the image, verifying if either the head entity or tail entity corresponds to the label of the object. We increased the knowledge base from 3,400 triples to a substantial 99,586 triples. With this substantial increase, extracting relevant knowledge from the knowledge base and answering the questions would be challenging. Due to the large size of the question-answering dataset and computational constraints, we primarily conducted experiments when filtering context is based on both images and questions. The results of our approach and previous baselines are presented in Table 3. Our model achieves an accuracy of 85.80% which is **5.95%** better than SOTA (Gao et al., 2023). Observations from Table 2 and Table 3 indicate that providing a dynamic number of triples increases accuracy.

Due to space constraints, comprehensive training setup details required for reproducing the results, including hyperparameters, hardware information, computational time, etc., are detailed in

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 4). The inference is done while providing the dynamic number of triples as context.

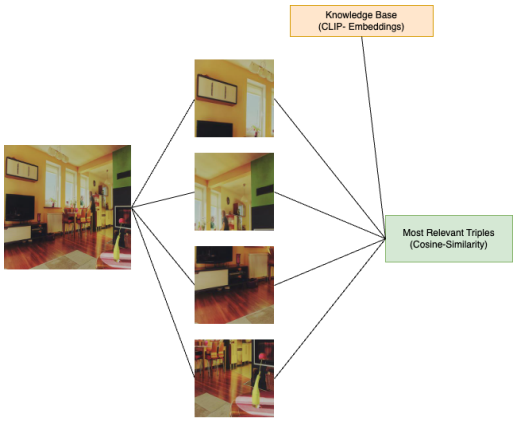


Figure 3: Splitting the image into four patches to extract relevant triples.

Appendix A.

5 Generalisation Capability

We showcase the generalization capability of our model by fine-tuning our pre-trained OFA model, initially trained on the KVQA dataset, on the FVQA dataset under various settings. 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 was 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 the Appendix D.

	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: Exact match scores with and without fine-tuning(FT, ∼FT) on the FVQA dataset. The model was pre-trained on the KVQA dataset. Context type (pre-training) denotes the type of context in which pre-training was done on the KVQA dataset. Context type (inference) denotes the type of context in which fine-tuning and inference were done on the FVQA dataset.

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. Consequently, valuable information from the image could be missed. The cosine similarity between embeddings of each patch and all the triples is then calculated to identify the most relevant triples. In our approach, we select the triples having a similarity greater than 0.8 for each patch. We further show the results without segmenting the image into patches and the results by identifying triples relevant to objects. For identifying triples relevant to objects, we utilised the object detection model, Detectron (Wu et al., 2019), to obtain bounding boxes for each object in the image. For each bounding box, we identified the most relevant triples. The results and detailed explanation of these approaches are presented in the Appendix C. For extracting the triples relevant to the question we use the same approach as explained in Section 3.1.2.

We use the same contextual settings as explained in Section 3.1.2. Table 6 displays the exact match scores obtained by the pre-trained OFA model, both with and without fine-tuning on the FVQA dataset. The context utilized during fine-tuning is identical to that used during inference. We observe that when providing the dynamic number of triples as context during pre-training or fine-tuning, the model demonstrates better reasoning capabilities compared to using a fixed number of triples as context.

Despite the FVQA and KVQA datasets belonging to different domains and the small size of the fine-tuning dataset (FVQA), our model demonstrates

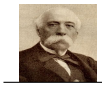
Image	Question	Truth Value	No Triples	All Triples	Filtered Triples
	Is the person in the image a politician?	No	Yes	No	No
	For how many years did the person in the image live?	83	72	82	83
	Were all the people in the image born in the same country?	No	Yes	Yes	No
	Who among the people in the image ever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

Table 7: Qualitative analysis, which presents instances from the dataset and their answer predictions with and without the presence of triples in the input.

superior generalization, achieving an accuracy of 65.28%, which is **4.18%** above the SOTA. The baselines of the FVQA dataset are provided in Table 5 and additional results in Appendix C.

6 Qualitative Analysis

We will showcase various instances illustrating how the integration of knowledge enhances the predictive capabilities of the OFA model. Concurrently, the efficient filtration of knowledge reduces noise, bolstering the model's reasoning capacity and leading to highly accurate predictions. We choose some samples as shown in Table 7. The first row shows that giving extra information helps the model make correct predictions for simple questions with straightforward answers. In this case, whether or not we filter the knowledge triples, the correct answer is obtained. The subsequent two rows (Row 2 and Row 3) highlight that when we provide all triples to the model, it receives irrelevant knowledge (noise), leading to incorrect predictions. However, when we filter the knowledge triples based on the questions, the model predicts the correct answer. These questions fall into complex categories, with Row 2 belonging to a single-hop subtraction question and Row 3 being a multi-hop boolean question, necessitating the removal of unnecessary information or noise.

In the final example, regardless of whether we provide knowledge triples 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 potential reason for the incorrect predictions is the limited number of questions in these complex categories, along with the model's relatively small size (472 million parameters). To perform such intricate inferences, larger models with billions of parameters may be necessary. Due to space constraints, we provide these and some more examples in the Appendix E Table 13.

7 Conclusion and Future Work

We introduced a novel approach for KBVQA utilizing a dynamic triple filtering module for retrieving external context, leveraging knowledge graphs. The proposed approach surpasses the SOTA on 3 different KBVQA datasets, on average by **4.75%**. We illustrated that furnishing the model with a dynamic number of triples as context during pretraining or fine-tuning leads to improved reasoning capabilities compared to utilizing a fixed number of triples. We also demonstrated the generalization prowess of our approach by achieving SOTA-beating performance on a small dataset, utilizing a model trained on data from an entirely different domain. We highlighted specific instances demonstrating how integrating relevant knowledge significantly improves our model's accuracy. Additional examples are available in the Appendix E. The main insight is that a dynamically determined number of relevant triples in the context eliminates noise, leading to more precise predictions.

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.

Limitations: Our research has specific limitations. Our approach requires calculating the similarity between the question and all the triples associated with the images to identify the most relevant ones. This process can become computationally intensive, especially when dealing with a substantial number of triples, resulting in longer prediction times.

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.

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 multimodal 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. Training was conducted on Nvidia RTX A6000 [1], with each iteration taking approximately 8 and 12 hours for the Base and Large models, respectively.

The number of parameters used and the number of encoder-decoder layers for both the OFA-Base and OFA-Large models is given in Table 9.

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 8. 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 10.

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.

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

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 11. 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 12.

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

Types of Questions	MEMNET	UNITIER	OFA(Ours) Single-Hop	OFA(Ours) 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 8: 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 9: 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	OFA(Ours) 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 10: 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").

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 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

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

Table 11: 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 12: 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.

to get relevant triples.

E More Examples

Refer to Table 13 for the examples used in Section 6, as well as some additional examples that demonstrate the effectiveness of our approach. Table 13 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.

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