Exploring varied sources of Commonsense Knowledge towards Commonsense VQA Tasks for small VLMs

Anonymous ACL submission

Abstract

Commonsense VQA tasks require the ability to reason about an image, the accompanying question, and commonsense knowledge not stated explicitly in the input text. Small Vision-Language models (VLMs) such as ViLT, VisualBERT still struggle with tasks which require knowledge about everyday objects, human interactions, and events. This necessitates the integration of external knowledge to enhance their reasoning capabilities. In this work, we perform a detailed analysis of the effect of integrating retrieved facts and generated explanations to smaller Vision-Language models (sVLM) on two popular commonsense QA datasets (CRIC, e-SNLI-VE). We benchmark various dense retrieval mechanisms, including finetuning ColBERTv2 (with small amounts of data) for retrieving commonsense facts. For prompting LLMs, we explore adding various contextual information such as, caption, objects, dense captions and retrieved facts. Interestingly, adding the groundtruth label in the presence of an incomplete context (such as a short caption) often leads to *comprehensive* but *hallucinated* explanations. Using dense captions and color information about objects, we generate both *informative* and *useful* explanations for vanilla sVLMs; which fare better than retrieved facts. Our results highlight the nuances of using LLMs as a commonsense knowledge base for visual QA.

1 Introduction

From the initial days of Deep Neural Network-based Visual Question-Answering (VQA) systems, many researchers have highlighted the need for external knowledge for VQA tasks (Wang et al., 2015; Aditya et al., 2018, 2019). While the benefit from accessing external knowledge for factoid VQA is well-explored (Wang et al., 2018, 2015; Marino et al., 2019; Chen et al., 2024), the benefit of the same for commonsense reasoning is under-explored in the context of small pretrained Vision-Language models (VLM). Previous explorations with commonsense knowledge integration (Aditya et al., 2018, 2019; Shah et al., 2019) did not involve pre-trained VLMs, apart from some recent ones (Ye et al., 2023). However, even recent explorations (Lin and Byrne, 2022; Yang et al., 2023) focus on factoid VQA tasks. Unlike encyclopedic knowledge which is used in open-domain Question-Answering (QA) tasks, commonsense knowledge required for Visual QA is diverse, ambiguous, and may span a vast number of topics. Therefore, the first challenge is finding suitable knowledge sources; a singular source may not always be sufficient. In this era of pre-trained Language models, such knowledge may come from commonsense knowledge graphs (KG), text corpora, or pretrained Large Language Models (LLM) which show a good extent of commonsense knowledge in many tasks (Ghosal et al., 2023). Therefore in this work, we raise and investigate two central questions: *1) to what extent can commonsense knowledge retrieved from external knowledge-bases, (KB) and LLM-generated explanations, help small VLMs? And 2) is one better suited than the other for commonsense reasoning in VQA systems?*.

We extensively analyze the effect of integrating retrieved commonsense facts (from existing commonsense knowledge bases) and LLM-generated explanations into several small VLMs. We select two generic architectures — a fully Transformers-based ViLT (Kim et al., 2021), and a separate image (CNN) and text encoder (BERT) based VisualBERT (Li et al., 2019). We first benchmark these models on two datasets CRIC (Gao et al.), and e-SNLI-VE (Xie et al., 2019). Our analysis shows the lack of commonsense knowledge in small *VLM*s to be the major cause of degraded performance. Using ablation studies with groundtruth explanations (present in above datasets), shows simply prepending appropriate knowledge can bring significant boost (upto 10%) to test accuracy. There-

fore, we first explore integrating retrieved facts from commonsense knowledge bases. Further, using a textual representation of the image and the question in the form of captions, list of objects and retrieved facts; we generate explanations using LLMs (Llama-3-8B)[1]. We observe, that if the context includes the groundtruth label, it helps Llama-3 generate explanations which results in high end-to-end test accuracy. Interestingly, qualitative analysis reveals hallucinations in such generated explanations. Therefore, we experiment with the effect of enriched context on LLM-generated explanations – which helps boost the performance significantly. Apart from concatenating explanations (or facts), we also explore simple architectural variations which are better equipped to reason with noisy retrieved knowledge or irrelevant explanations. We make the following contributions:

- We benchmark several dense retriever based mechanisms for retrieving *missing* commonsense facts necessary for visual commonsense QA. We obtain the most benefit from finetuning ColBERTv2 on a subset of data (utilizing groundtruth explanations as target from CRIC, e-SNLI-VE).
- We explore various ways to capture visual context in text and use that with LLMs (Llama-3-8B) to generate explanations. When Llama-3 is provided with the true label, we observe that simply prepending such explanation to the context achieves a much higher boost in accuracy over vanilla ViLT/VisualBERT, compared to the same with groundtruth explanations (in CRIC, e-SNLI-VE). Further analysis shows that groundtruth label makes the LLMs efficient post-hoc explainers, while incomplete context (such as the caption) can promote *halluncinations*. Using dense captions from a VLM (Florence-2-large) to capture the context achieves a better tradeoff.
- We perform additional ablations with various types of retrievers, different types of queries, and LLMs for generating explanations.

2 Related Work

Knowledge based VQA systems poses a challenging task, as it needs the models to comprehend visual and textual data while utilizing external knowledge to derive correct answers. For instance as depicted in the Fig. 1, "is there a place that is blue and liquid?". The correct result "Ocean" has to be deduced using the visual information, common knowledge that it is "liquid" and "blue". Symbolic structured knowledge is helpful in this context, for instance existing literature has used knowledge graphs to retrieve facts, using knowledge graph embeddings and formal language queries (Wang et al., 2018; Zheng et al., 2021; Chen et al., 2021). For unstructured knowledge, Karpukhin et al. (2020) showed that the Dense-Passage-Retrieval (DPR) technique perform the best by encoding both the question and the knowledge sentences to text embedding using BERT. This method uses cosine similarity (or the dot product) between vectors to retrieve relevant knowledge text.

Retrieval-Augmentation for sVLM. Learning with retrieval-augmented information has been widely used in language modelling research. In recent literature, retrieval-augmented frameworks for joint vision-language and other multimodal modelling tasks have also been studied more thoroughly. Retrieval-based augmentation methods are applied in a few different ways in recent work:

i) Retrieval based on images: Rao et al. (2024) proposed a technique to retrieve image-caption pairs given the input image with a CLIP (Radford et al., 2021) based scoring technique. The retrieved image-caption pairs and the given input image with the query are then used by an OFA (Wang et al., 2022) model (with 182M parameters) as the VLM to generate the task-specific textual output. Wang et al. (2023) also proposed a method to retrieve similar images and incorporate them as part of vision-language attention layers in a modified transformer-based language decoder model. Rao et al. (2023) presented a method to retrieve knowledge graphs (KG) based on the given input image. The retrieved KG is then encoded with a graph encoder and used as part of a knowledge-augmented vision-language model with BERT (Devlin et al., 2019) and ViT-B (Dosovitskiy et al., 2021). These works consider VQA and image captioning as the end tasks and show promising results by incorporating the retrieval mechanism.

ii) Retrieval based on multimodal content: Hu et al. (2023) and Yasunaga et al. (2023) both proposed techniques to perform retrieval based on multimodal input to obtain multimodal outputs and use them in their language modelling framework. Hu et al. (2023) also used an end-to-end training tech-

[1]In this work, we purposefully limit ourselves to "small" VLMs. So, we do not use very large LMs such as `gpt-3.5` for explanation generation.

nique with T5 (Raffel et al., 2020) as the language model and ViT (Dosovitskiy et al., 2021) as the image encoder totalling a parameter count between 400M - 2.1B in their model variants. The end-to-end method shows improved performance over relevant baselines in the AOKVQA dataset. The method of Yasunaga et al. (2023) can also generate images and they report promising performance in both image-to-caption and caption-to-image generation tasks. Caffagni et al. (2024) proposed a two-stage retrieval framework that uses the input image in the first stage to retrieve relevant documents from Wikipedia using CLIP; and then uses the input question in the second stage to retrieve relevant passages from those documents using Contriever (Izacard et al., 2021). Finally, the retrieved passages are used in the LLaVA family of VLMs (having more than 7B parameters in the language model) to perform visual question answering. The second stage of the retrieval based on the input question is close to our proposed techniques in this paper. However, we exploit additional textual information for retrieval; retrieve from sources other than Wikipedia; and use the retrieved information in a much smaller VLM.

3 Problem Definition: Role of External Commonsense Knowledge in Commonsense Visual QA

In Visual Question Answering (VQA) datasets, often questions inquire about specific details of the image, necessitating an understanding beyond factual information perceptible from the image. For instance, when presented with an image of a group of giraffes, a question may ask, *"What is the object that has the head and is a type of even-toed ungulate doing?"* The answer to such a question requires the model to possess implicit commonsense knowledge, recognizing that a giraffe is a type of "*even-toed ungulate*." Similarly, in an indoor scene, a question like "Is there a gray furniture that can be used for sitting on?" entails the understanding that certain objects, like sofas or chairs, are a type of "furniture" and are used for sitting. Additionally, commonsense datasets may naturally introduce ambiguities due to existence of multiple correct answers, resulting in model's inability to to discern implicit commonsense knowledge. This hinders the system's ability to navigate subtle contextual nuances, impacting the reliability of answers provided for intricate queries about images. Addressing these limitations is pivotal for enhancing the contextual understanding of VQA systems.

Here, we lay down the task more formally. The input for our task comprises an image (I) and a natural language question (Q), and the objective is to answer the question while utilizing a set of implicitly required commonsense knowledge fact(s) (K'). We utilize the ground-truth answer (A) and the ground-truth explanation/knowledge (f^*), when available, as supervision.

4 Knowledge-Augmented Visual QA framework

Here, we explore two types of knowledge sources: traditional commonsense knowledge corpora (textual versions), and pretrained Large Language Models. Following Chen et al. (2024), we can represent a retrieval or knowledge augmented VQA system probabilistically as follows:

$$p(A|I,Q,\mathcal{K}) = q_\Theta(\mathcal{F}_{ret}|Q,I,\mathcal{K})p_\Phi(A|Q,I,\mathcal{F}_{ret}),$$

where, $q_\Theta(\cdot)$ is termed as retriever, which is tuned to retrieved top k relevant facts ($\mathcal{F}_{ret}$), and $p_\Phi(\cdot)$ is termed as reader, which utilizes the question, image and retrieved facts to predict the answer. To generalize to LLM-generated explanations, we can think of the reader as a model which outputs a paragraph relevant to the question. We will utilize the following notations wherever required:

$$p(A|I,Q,\mathcal{M}) = q_\Theta(\mathcal{E}_{ret}|Q,I,\mathcal{M})p_\Phi(A|Q,I,\mathcal{E}_{ret}),$$

where $\mathcal{M}$ may stand for a knowledge base or an LLM. $\mathcal{E}_{ret}$ may stand for a set of retrieved facts and LLM-generated explanation(s).

4.1 Retrieving Knowledge from Commonsense Corpus

Commonsense Corpus. Commonsense knowledge includes basic facts and anecdotes about everyday objects, actions, activities; commonly assumed by humans to make inferences about everyday observations (Li et al., 2022). As current Transformer models seem to reason well with commonsense facts and rules written in text (Clark et al., 2021), we utilize commonsense in natural language form that is more varied and nuanced, compared to the existing structured knowledge graphs. Yu et al. (2022) introduced a large, diverse natural language-based commonsense corpus consisting of 20M commonsense knowledge facts from multiple human-annotated commonsense knowledge bases,

commonsense-related benchmarks datasets, and other commonsense-related data scraped from the web. Based on initial ablations (see Appendix Tables 5 and 6), we chose only the subset of facts obtained from the Open Mind Commonsense (OMCS; Havasi et al. (2010)) corpus due to their significantly higher similarity scores with the ground truth explanation in our target datasets and smaller corpus size. OMCS contains $\sim$ 1.5M natural language sentences (facts) exhibiting commonsense information about everyday events and objects.

Commonsense Knowledge Retrieval. For a given input sample (image-question pair), the goal is to retrieve a set of k knowledge facts (where each fact $f_{i \leq k}$ is a sentence or phrase) from the commonsense corpus ($\mathcal{K}$), such that the facts provide relevant commonsense knowledge about the input. Ideally, retrieval requires abductive reasoning or finding out the "missing pieces" of information required for performing certain deductions. We explore primarily text-based dense retrieval mechanisms utilizing SBERT, ColBERTv2, and the recently introduced stella-en-v5. We represent the query in the following ways:

- **Question**: We use the question as the query.
- **Caption + Question**: We generated image captions using BLIP-2 (Li et al., 2023). We prepend the caption to the question to form the query.
- **Objects + Question**: We use a pre-trained Yolov8 (Redmon et al., 2016) to detect objects from the image. We use the list of detected objects along with the question as the query.
- **Scene Graph Text + Question**: We generated scene graph from images using Relation Transformers (Cong et al., 2023). We prepend the sequence of scene graph triplets to the question as the query.

We use SBERT (Reimers and Gurevych, 2019) to create embeddings of the entire commonsense corpus ($\mathcal{K}$) which were then indexed using FAISS (Douze et al., 2024). We then use the same encoder to encode each query (for example, question/hypothesis) and perform Nearest Neighbour Search using the indexed corpus embeddings to retrieve the top-k closest facts. Provided SBERT captures semantic similarity, to improve retrieval results we use pretrained ColBERTv2, which is specifically trained with the retrieval objective. The improved performance (evident from Table 1) motivated us to further finetune ColBERTv2 for commonsense facts retrieval.

Figure 1: Facts retrieved by pre-trained ColBERTv2 vs. Finetuned ColBERTv2 vs. FAISS. Facts retrieved using pre-trained ColBERTv2 are more semantically close and contextually relevant to the query compared to the FAISS similarity search over SBERT embeddings. Although the facts retrieved by the pre-trained model are useful, they are not quite relevant to the question which asks for a *place*. This shows that finetuning the retriever resulted in retrieved facts relevant to the image along with the query supplied.

Finetuning ColBERTv2. As shown in Fig. 1, pre-trained ColBERTv2 can retrieve partially relevant facts as they are pre-trained on MS-MARCO and TREC-CAR which are generalist datasets. Therefore, following Khattab and Zaharia (2020), we finetune ColBERTv2 by preparing the commonsense corpus in the format of triples $\langle q, d^+, d^- \rangle$ where q is the query, d^+ is the ground-truth knowledge, and d^- is a randomly sampled fact from the corpus. It employs a contrastive learning approach by computing the pairwise softmax cross-entropy loss over the relevance score $S_{q,d}$ computed independently for d^+ and d^- with the query q.

$$L(q, d^+, d^-) = -\log \left(\frac{\exp(S_{q,d^+})}{\exp(S_{q,d^+}) + \exp(S_{q,d^-})} \right)$$

We used a batch size of 32, the maximum length of the facts as 128 (most facts are much smaller in length), and used default settings for other parameters as in Khattab and Zaharia (2020). We used $160K$ samples from the training splits of both CRIC and e-SNLI-VE and finetuned first on CRIC, then on e-SNLI-VE.

Commonsense Knowledge Integration. The retriever-reader architectures can be trained in two

ways: 1) in an end-to-end fashion where the retriever and the VLM are finetuned together (Sachan et al., 2024), and 2) where they are finetuned independently of each other. For simplicity, we go with the latter approach to improve both independently as 1) this allows for a modular plug-and-play framework wherein the retriever/reader can be switched as per the specific task. 2) In the case of sub-optimal retrieval performance, the reader ($p_\Phi(\cdot)$) is expected to learn to reason with noisy added knowledge and selectively attend to only the relevant parts of the added knowledge with respect to the image and question/hypothesis.

4.2 Integrating LLM-generated Explanations

Using the image I, we extract various pieces of information such as multiple types of captions, list of objects present, and set of retrieved facts to textually represent the visual context. We use the following prompt for the instruction-finetuned LLMs (such as Llama-3) to generate an explanation[2].

> Given an image description, the task is to generate an explanation based on the following information.
> ① Caption: $\texttt{caption}_I$.
> Dense caption: $\texttt{densecaption}_I$.
> Region Captions: $\texttt{regioncaption}_I$.
> Objects Present: $\texttt{Objects}_I$
> ② Question: Q.
> ③ Retrieved Facts are: $F_{I,Q}$
> ④ Ground-truth label: $y_{I,Q}$
> You strictly need to produce 15-20 words single line explanation which can help VQA models derive the conclusion and nothing else. The forbidden words are: "image description", "captions".

As shown above, we divide the information provided in four groups. Other than the post-hoc setting, we do not use ④ the ground-truth label. Interestingly, Llama-3-8B produces high-quality explanations (silver truth) when provided the ground-truth label. Manual analysis shows such explanation heavily suffers from hallucinations when the captions do not capture question-relevant information. Therefore, we also perform detailed analysis of the effect of varying ① information on generating explanation. The variants are of the following:

- **Type 0**: Caption, ② + ③ + ④.
- **Type 1**: Caption, ② + ③.
- **Type 2**: Dense Caption, ② + ③.
- **Type 3**: Region Caption, ② + ③.
- **Type 4**: Region Caption + Objects, ② + ③.
- **Type 5**: Dense Caption + Region Caption + Objects, ②+③.
- **Type 6**: Dense Caption + Region Caption + Objects, ② (Type 5 without knowledge facts).
- **Type 7**: Florence Finetuned Explanations

Integrating Explanations. Since explanations are more targeted and concise, we explore baseline variations by simply prepending the explanation with the question for the small VLMs.

5 Experimental Setup

5.1 Datasets

VQA tasks lie at the intersection of vision, language, and commonsense reasoning. The task requires the model to produce an answer given an image and a question in natural language. The annotated answers are originally in free-form natural language, but converting it to multiple-choice format for evaluation is common practice. There exists established benchmark VQA datasets such as VQAv2 (Goyal et al., 2017), VCR (Zellers et al., 2019), VisDial (Das et al., 2017), etc. In this work, we choose two popular VQA datasets that require commonsense reasoning: CRIC (Gao et al.) and e-SNLI-VE (Xie et al., 2019) due to their size, diversity of questions, and availability of gold truth explanations.

CRIC provides 494K automatically generated questions on 96K images from Visual Genome along with 3.4K knowledge items from ConceptNet and Wikipedia. Each sample is formed as a multiple-choice question with 4 possible answer options of which only one is correct. e-SNLI-VE is based on two existing datasets– Flickr30k and SNLI-VE. Each image in Flickr30k is used as the visual context. For each image, the authors collect multiple hypotheses that can be entailed or contradicted by or is neutral with respect to the image, resulting in about 400K (image, text) pairs.

5.2 Models

We choose two baseline Vision-Language models – VisualBERT (Li et al., 2019) and ViLT (Kim et al., 2021) – spanning two broad modality fusion architectures.

[2]The prompt mentioned here is used for Llama-3-8B. The GPT3.5 prompt is mentioned in the appendix.

ViLT is a Transformer-based unified architecture that uses a single transformer module to encode and fuse both visual and text features in place of a separate deep visual embedder. This design inherently leads to significant runtime and parameter efficiency. ViLT processes both modalities in a unified manner through self-attention layers.

VisualBERT requires a set of visual embeddings (representing the visual features) extracted from an image encoder and text embeddings encoded with BERT. Then, it utilizes the self-attention mechanism for implicit alignment of elements in the input text and regions within the input image.

6 Results

6.1 Baseline Performance

To establish an estimate of the vanilla performance (without any knowledge addition), we trained ViLT and VisualBERT with only the image and the question/hypothesis on CRIC and e-SNLI-VE. This yielded an accuracy of 74.73% on CRIC and 76.60% on e-SNLI-VE respectively. Our analysis shows that ViLT faces specific challenges on colour-based and commonsense-based questions. We also find that the train set of CRIC has a non-negligible amount of label noise, which may cause the models to learn erroneous patterns in data.

6.2 Analysis of Standalone Retrieval Performance

We first perform a detailed ablation over various retrieval methods to choose the best retriver system. In Table 1, we show our retrieval results on CRIC test split by comparing the retrieved fact(s) with the provided groundtruth explanations. Because FAISS uses a static index produced from the SBERT embeddings and employs simplistic measures of similarity like geometric distance or cosine similarity, it cannot retrieve *missing* pieces of knowledge required for abductive reasoning with respect to the query. This is demonstrated by Fig. 1 where the facts retrieved by FAISS are conceptually close to the question but not relevant, whereas the facts retrieved by ColBERTv2 are exact answers. Table 1 further shows that utilizing pre-trained ColBERTv2 (Khattab and Zaharia, 2020) embeddings significantly outperforms FAISS in terms of BLEU, ROUGE, and Cosine similarity scores across various query settings. For instance, when $k = 5$, ColBERTv2 achieves a BLEU-1 score of 56.24, significantly higher than FAISS's best score of 38.30 with the same query setup. Similarly, the ROUGE-L score of ColBERTv2 at $k = 10$ reaches 67.32, compared to FAISS's 47.80. The examples in Fig. 1 together show how finetuned ColBERTv2 learns to retrieve more relevant commonsense knowledge.

6.3 Capturing Visual Context for LLM-Generated Explanations

In Table 2, we show the comparison of various prompt settings and the effects of different information in generating the explanation. Our results on $20k$ datapoints from CRIC validation split shows that **Type 5** variant with Dense captions, region captions and objects outperform other types. Clearly, region captions add more relevant information. Our previous analysis of baseline has shown that commonsense questions asking about colors of objects are the hardest for small VLMs. As, region captions and dense captions detect the color, the explanations are more relevant. **Type 5** also outperforms variants without retrieved facts – thereby showing the importance of facts retrieval while generating explanations.

6.4 Appending LLM-generated Explanations with groundtruth label(s)

From Table 3, it can be observed that the generated explanations (with the label) consistently outperform the ground truth explanations by a significant margin for all architectures and datasets. The margin is up to 20% for CRIC when using VisualBERT, and roughly 6% for e-SNLI-VE on ViLT. However, the explanations generated without the label fare quite poorly and the performance is much lower than even the vanilla baseline performance. We conduct a qualitative analysis of a subset of the **Type 0** and **Type 1** explanations to understand what allows the with-label generated explanations to enable near-perfect performance while the ones generated without the label degrade performance so significantly. We see that often providing the labels forcefully *rectify* for missing context (see Fig. 3(b)). In many cases, the explanations are written in a more assertive manner than the groundtruth, such as Fig. 3(d) **Type 0** states "the maccaroni is the main course..." than "Macaroni is usually treated as the main course.". Our analysis also shows that **Type 0** often results in hallucinated explanations, by somehow "connecting" the context and the answer. For **Type 1** explanations, such benefits disappear. Our analysis on 400 examples from CRIC shows, **Type 1** explanations are am-

Query	Method	R-L@5	R-L@10	B-1@5	Cosine@5	Cosine@10
Q	FAISS	41.86	47.80	38.30	53.56	58.46
C + Q	FAISS	38.57	44.70	35.21	51.54	56.88
Objects + Q	FAISS	40.93	46.73	37.34	52.61	57.67
SG + Q	FAISS	36.05	41.91	32.88	47.01	52.21
All + Q	FAISS	35.52	46.55	32.65	46.55	51.78
Q	ColBERTv2	61.33	67.32	56.24	68.69	73.11
Q	ColBERTv2-FT	**74.48**	**77.44**	**69.99**	**77.65**	**80.62**
All + Q	ColBERTv2-FT	19.40	23.39	16.94	33.89	37.67
Q	stella-en-v5	46.72	54.27	42.01	62.58	67.18
Objects + Q	stella-en-v5	43.89	51.36	39.27	62.98	67.66
All + Q	stella-en-v5	29.01	34.65	25.67	46.76	51.15

Table 1: Comparison of BLEU, ROUGE, and Cosine Scores for Various Retrieval Mechanisms (on CRIC test split). We report the performance of different retrieval methods based on BLEU-1 (B-1), ROUGE-L (R-L), and Cosine similarity scores across different query settings. The results are presented for top-5 and top-10 retrieved facts. Finetuned ColBERTv2 outperforms all other methods. Q: Question, C: Caption, and SG: Scene Graph.

Type	Explanation	B-1	B-2	B-3	R-1	R-2	R-L	Cosine
Type-0	C + Q + $F_{I,Q}$ + $y_{I,Q}$	18.78	11.58	8.46	36.35	13.21	32.36	53.54
Type-1	C + Q + $F_{I,Q}$	24.97	16.92	13.06	40.40	18.08	37.14	55.09
Type-2	DC + Q + $F_{I,Q}$	27.42	19.25	15.22	42.82	20.78	39.41	58.75
Type-3	RC + Q + $F_{I,Q}$	31.34	23.04	18.62	46.55	25.12	43.43	60.22
Type-4	RC + O + Q + $F_{I,Q}$	27.47	20.72	16.84	46.71	**25.35**	**43.48**	60.26
Type-5	DC + RC + O + Q + $F_{I,Q}$	**31.45**	**23.10**	**18.71**	**46.79**	25.09	43.29	**60.70**
Type-6	DC + RC + O + Q	31.27	22.77	18.15	45.59	23.93	42.31	59.10
Type-7	I + Q + <*TP*>	27.43	20.45	16.42	45.59	23.93	42.31	59.07

Table 2: Results from a 20K subset from the val set of CRIC showing a comparison of BLEU, ROUGE, and Cosine Scores (w.r.t the GT explanation) for various explanation generation mechanisms. We report the quality metrics of explanations based on BLEU-1 (B-1), BLEU-2 (B-2), BLEU-3 (B-3), ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L), and Cosine similarity scores across different settings. Q indicates Question, O indicates Objects, C indicates Caption, $F_{I,Q}$ indicates Retrieved Facts, DC indicates Dense Caption, RC indicates Region Caption, I indicates Image and $< TP >$ indicates task prompt.

	Architecture	ViLT	VisualBERT
e-SNLI	Ground truth Expl.	92.47	93.49
	LLM Expl. (**Type 0**)	**97.91**	**97.89**
	LLM Expl. (**Type 1**)	75.92	76.11
CRIC	Ground truth Expl.	82.98	77
	LLM Expl. (**Type 0**)	**97.02**	**96.60**
	LLM Expl. (**Type 1**)	55.65	53.82

Table 3: Performance (accuracy) comparison between ground-truth explanations and generated explanations (with and w/o) labels on e-SNLI and CRIC.(**Type 0**): Post-hoc with GT label as input.

biguous 78.6% times. **Type 0 (Oracle) vs Type 5 Explanations.** We analysed 100 randomly selected samples from the CRIC test set to evaluate the relevance and hallucinations in **Type-5** vs. **Type-0** explanations. We found that around 45% of the **Type-0** explanations are hallucinated. Due to the the gold label in the prompt, the LLM attempts to align the generated explanation with the given label even if the *explanation is spurious*. Among **Type-5** explanations, we found that only around 20% of the explanations have irrelevant content. As this setting uses information from the dense captions (DC) and region captions (RC), more than 75% of the explanation contains relevant information to derive the answer due to better scene understanding. We also finetuned Florence-large for explanation generation bypassing the LLM's whose performance is reported in 2. We illustrate some of the florence generated explanations in appendix.

6.5 End to End Performance Analysis

Integrating external knowledge from external sources to enhance multimodal capabilities has been an area of interest for past few years. Here, we perform extensive experiments on the effects of retrieved facts and generated explanations on small Vision-Language models. We integrate knowledge

Architecture	Retrieved/ Generated	Accuracy	
		e-SNLI-VE	CRIC
ViLT	✗	76.46	72.99
Explanation-Gating (ViLT)	**Type 1**	76.91	63.04
Concat (ViLT)	**Type 5**	**78.46**	71.76
Concat (ViLT)	**Type 7**	76.44	71.24
Explanation-Gating (ViLT)	**Type 5**	*77.60*	65.04
Majority Voting (ViLT)	*Q/CB-FT*	74.51	69.08
Cross-Att (ViLT)	*Q/CB-FT*	76.39	**74.39**
VisualBERT	✗	74.48	62.60
Concat(VisualBERT)	**Type 5**	**78.83**	**69.69**
Concat(VisualBERT)	**Type 7**	**76.88**	**66.72**
Majority Voting (VisualBERT)	*Q/CB-FT*	72.53	60.25

Table 4: Performance of various external knowledge integration architectures on CRIC and e-SNLI-VE. *Q/CB-FT*: Facts retrieved by fine-tuned ColBERTv2 with question as the query.

with the vanilla baselines using the following variations:

- **Concat**: Here we prepend the explanation (or facts) to the context, as an input to the model.
- **Explanation-Gating**: We computed a relevance score between the explanation ($\mathbf{x_e}$) and the hypothesis/question embedding ($\mathbf{x_q}$) as a dot product. We then multiply the explanation hidden state with computed relevance score ($\tilde{\mathbf{x}}_\mathbf{e} = \mathbf{x_e} \times \mathbf{x_e}^\top \mathbf{x_q}$). We concatenate $\tilde{\mathbf{x}}_\mathbf{e}$ with the hypothesis/question embedding and then feed the sequence to a classifier. The goal is to weigh the importance of the explanations with respect to the hypothesis/question.
- **Majority Voting** (Retrieval): Following recent literature, we train the model with prepended top-1 fact, similar to **Concat**. During inference, we use top-k facts one by one to infer the label and perform majority voting.
- **Cross-Att**: This module utilizes multi-head self-attention to perform cross-attention between the encoded hypothesis and explanation, capturing their contextual interactions.(Refer to Section A in the Appendix).

We observe that our detailed analysis of the role of context for explanation generation helps the most. Simply prepending the **Type 5** explanations provide a 4.4% and 7.09% increase for VisualBERT on e-SNLI-VE and CRIC. The improvement for ViLT is low, around 2%.

7 Conclusion

We investigated whether adding knowledge from external knowledge sources can help small Vision-Language Models (sVLM) perform better on commonsense VQA tasks. We particularly focus on identifying *which* source of knowledge – whether from textual commonsense corpora or Large Language Models –provides the greatest benefit. Among dense retrievers, we observe finetuning ColBERTv2 on commonsense VQA dataset becomes the most effective retrieval strategy, even surpassing recent pretrained embedding methods such as `stell-en-v5` (one of the state-of-the-art strategies on popular retrieval benchmarks). For generating useful and relevant explanations with LLMs, we benchmark various ways to capture visual context correctly. We observe the use of dense captions and region captions generated using pretrained Florence2 offers the most benefit. Interestingly, using retrieved facts as additional context, helps as well. When prepended with the context, such explanations provide end-to-end accuracy boosts upto 7.09% for VLMs such as VisualBERT. Our research shows using LLMs as knowledge source can be an effective strategy and hallucination and irrelevant content can be diminished by providing detailed dense captions and relevant retrieved facts.

Limitations

In this work, we do not observe the effect of better knowledge integration architectures such as KAT (Gui et al., 2022), RA-VQA (Lin and Byrne, 2022), ReViLM. Our primary goal was to delve deeper into suitable knowledge sources and how to elicit useful and relevant explanations from Large Language Models.

At present, we restrict ourselves to `Llama-3-8B`. Our initial experiments with `Llama-2` did not produce better results. Our results show that we can significantly bring down hallucinations and increase relevance by representing the visual context with informative captions, and supplyinf relevant facts from commonsense knowledge bases. Currently, we can not claim that all Large Language Models will behave similarly as `Llama-3-8B`. We wish to investigate this as a separate phenomena in the future, and therefore, believe this is out of scope of the current work.

References

Somak Aditya, Yezhou Yang, and Chitta Baral. 2018. Explicit reasoning over end-to-end neural architectures for visual question answering. In *Proceedings*

of the AAAI Conference on Artificial Intelligence, volume 32.

Somak Aditya, Yezhou Yang, and Chitta Baral. 2019. Integrating knowledge and reasoning in image understanding. In *28th International Joint Conference on Artificial Intelligence, IJCAI 2019*, pages 6252–6259. International Joint Conferences on Artificial Intelligence.

Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1818–1826.

Zhuo Chen, Jiaoyan Chen, Yuxia Geng, Jeff Z. Pan, Zonggang Yuan, and Huajun Chen. 2021. Zero-shot visual question answering using knowledge graph. *Preprint*, arXiv:2107.05348.

Zhuo Chen, Yichi Zhang, Yin Fang, Yuxia Geng, Lingbing Guo, Xiang Chen, Qian Li, Wen Zhang, Jiaoyan Chen, Yushan Zhu, Jiaqi Li, Xiaoze Liu, Jeff Z. Pan, Ningyu Zhang, and Huajun Chen. 2024. Knowledge graphs meet multi-modal learning: A comprehensive survey. *CoRR*, abs/2402.05391.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2021. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. 2023. Reltr: Relation transformer for scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Difei Gao, Ruiping Wang, Shiguang Shan, and Xilin Chen. Cric: A vqa dataset for compositional reasoning on vision and commonsense. *IEEE Transactions on Pattern Analysis and Machine Intelligence, year=2023, volume=45, number=5, pages=5561-5578, doi=10.1109/TPAMI.2022.3210780*.

Deepanway Ghosal, Navonil Majumder, Roy Lee, Rada Mihalcea, and Soujanya Poria. 2023. Language guided visual question answering: Elevate your multimodal language model using knowledge-enriched prompts. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12096–12102, Singapore. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *Preprint*, arXiv:1612.00837.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander G Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. Kat: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 956–968.

Catherine Havasi, Robert Speer, Kenneth Arnold, Henry Lieberman, Jason Alonso, and Jesse Moeller. 2010. Open mind common sense: crowd-sourcing for common sense. In *Proceedings of the 2nd AAAI Conference on Collaboratively-Built Knowledge Sources and Artificial Intelligence*, AAAIWS'10-02, page 53. AAAI Press.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. 2023. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23369–23379.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. *Preprint*, arXiv:2004.04906.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. *Preprint*, arXiv:2004.12832.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A survey on retrieval-augmented text generation. *Preprint*, arXiv:2202.01110.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *Preprint*, arXiv:1908.03557.

Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. 2018. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. *ECCV*.

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11238–11254, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Jiahua Rao, Zifei Shan, Longpo Liu, Yao Zhou, and Yuedong Yang. 2023. Retrieval-based knowledge augmented vision language pre-training. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5399–5409.

Varun Nagaraj Rao, Siddharth Choudhary, Aditya Deshpande, Ravi Kumar Satzoda, and Srikar Appalaraju. 2024. Raven: Multitask retrieval augmented vision-language learning. *arXiv preprint arXiv:2406.19150*.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. *Preprint*, arXiv:1506.02640.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Devendra Singh Sachan, Siva Reddy, William Hamilton, Chris Dyer, and Dani Yogatama. 2024. End-to-end training of multi-document reader and retriever for open-domain question answering. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

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.

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. 2015. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*.

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. In *International conference on machine learning*, pages 23318–23340. PMLR.

Weizhi Wang, Li Dong, Hao Cheng, Haoyu Song, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023. Visually-augmented language modeling. In *The Eleventh International Conference on Learning Representations*.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, Mohammad Shoeybi, Ming-Yu Liu, Yuke Zhu, Bryan Catanzaro, Chaowei Xiao, and Anima Anandkumar. 2023. Re-ViLM: Retrieval-augmented visual language model for zero and few-shot image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11844–11857, Singapore. Association for Computational Linguistics.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. 2023. Retrieval-augmented multimodal language modeling. In *International Conference on Machine Learning*, pages 39755–39769. PMLR.

Shuquan Ye, Yujia Xie, Dongdong Chen, Yichong Xu, Lu Yuan, Chenguang Zhu, and Jing Liao. 2023. Improving commonsense in vision-language models via knowledge graph riddles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2634–2645.

Wenhao Yu, Chenguang Zhu, Zhihan Zhang, Shuohang Wang, Zhuosheng Zhang, Yuwei Fang, and Meng Jiang. 2022. Retrieval augmentation for commonsense reasoning: A unified approach. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4364–4377, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wenfeng Zheng, Lirong Yin, Xiaobing Chen, Zhiyang Ma, Shan Liu, and Bo Yang. 2021. Knowledge base graph embedding module design for visual question answering model. *Pattern Recognition*, 120:108153.

Appendix

A Cross Attention Architecture Variant

We designed a cross-modal classification architecture to combine visual and textual information for image-question-answering tasks. The model uses a ViLT encoder to jointly encode image and hypothesis pairs, while an SBERT encoder processes the hypothesis and explanations separately. We apply cross-attention between the encoded hypothesis and explanation to capture interactions between them. The output of the cross-attention mechanism is pooled and fused with the ViLT encoder's output, after which the fused features are passed through a fully connected classifier to predict the final label. This architecture aims to leverage both visual and textual cues to enhance multimodal reasoning and classification.

$$
\begin{aligned}
\mathbf{v}_{\text{ViLT}} &= \text{ViLT}(\mathbf{I}, \mathbf{H}) \\
\mathbf{h} &= \text{SBERT}(\mathbf{H}), \quad \mathbf{e} = \text{SBERT}(\mathbf{E}) \\
\mathbf{v}_{\text{proj}} &= \mathbf{W}_v \mathbf{v}_{\text{ViLT}}, \quad \mathbf{h}_{\text{proj}} = \mathbf{W}_h \mathbf{h}, \quad \mathbf{e}_{\text{proj}} = \mathbf{W}_e \mathbf{e} \\
\mathbf{A} &= \text{CrossAttention}(\mathbf{h}_{\text{proj}}, \mathbf{e}_{\text{proj}}, \mathbf{e}_{\text{proj}}) \\
\mathbf{A}_{\text{pooled}} &= \max(\mathbf{A}, \text{dim} = 1) \\
\mathbf{f}_{\text{fused}} &= \text{Concat}(\text{mean}(\mathbf{v}_{\text{proj}}, \text{dim} = 1), \mathbf{A}_{\text{pooled}}) \\
\mathbf{f}_{\text{fused}} &= \mathbf{W}_{\text{fusion}} \mathbf{f}_{\text{fused}} \\
\text{logits} &= \text{Classifier}(\mathbf{f}_{\text{fused}})
\end{aligned}
$$

B Qualitative Analysis of Generated, Retrieved and Manual Explanations.

In Fig. 3, we show groundtruth, retrieved facts, and generated explanations (Type 0, 1 and 5) comparatively for images in CRIC. Ground-truth explanations are accurate but lack context grounded in the query making them less helpful for complex queries. **Type-1** generated explanations depend on the often incomplete context provided by the caption. For example, for Fig3 (b)), the caption is "there are many different types of food on display on the case" and the objects include only "bananas", as the orange is obscured from view by the box in front. The LLM hallucinates and claims that the banana is in behind the box whereas it is actually in front. But the **Type-5** explanation which includes region information correctly detects the orange behind the box and reasons that an *orange* has a higher vitamin C content and is also behind the box. Retrieved facts often consist of fragmented information that lacks the necessary context or specificity for the exact question. These facts may score higher on metrics like BLEU or ROUGE due to shared phrases with ground-truth explanations, but they often fail to contribute significantly to the model's reasoning process, as they do not contain the elaborate context required for accurate answer derivation.

C Comparison of Retrieval Performance from OMCS 1.5M and 20M Knowledge Corpus

To verify the commonsense facts quality we chose a random sample of 20K from both CRIC and e-SNLIVE dataset. Although the CRIC test set is larger, a subset of 20K facts was chosen to make the results comparable to that of e-SNLI-VE which has 15K samples in the test set. Looking at the results it is evident that the OMCS corpus alone contains more relevant facts for commonsense reasoning when compared to the Ground Truth explanations. OMCS is also nearly 20X smaller, hence retrievals from the corpus are significantly faster.

D Additional Examples of Retrieved Facts: FAISS vs ColBERTv2

We provide more examples in Fig. 2, depicting the difference in relevance of the facts retrieved by FAISS and those retrieved by ColBERT. In (a), both FAISS and ColBERT retrieve facts unrelated to the query which asks about *furniture*, whereas

Corpus	K	Rouge1	Rouge2	RougeL	BLEU	Cosine Score
OMCS 1.5M	@ 5	55.86	45.52	54.94	57.68	60.58
	@ 10	58.86	49.22	58.09	66.74	62.38
20M Comm. Corpus	@ 5	41.06	28.11	39.69	51.49	45.63
	@ 10	38.96	25.48	37.56	61.19	43.74

Table 5: BLEU, ROUGE, and Cosine Score of facts retrieved from the OMCS Corpus against facts retrieved from the cumulative 20M corpus introduced by (Yu et al., 2022) using pre-trained ColBERTv2. Retrieved facts are compared to the ground truth explanation. Scores reported on 20K random samples from the test set of e-SNLI-VE.

Corpus	K	Rouge1	Rouge2	RougeL	BLEU	Cosine Score
OMCS 1.5M	@ 5	28.58	13.71	25.86	48.12	30.67
	@ 10	28.52	13.71	25.77	47.96	15.37
20M Comm. Corpus	@ 5	27.84	12.50	24.81	41.63	26.21
	@ 10	27.63	12.49	24.95	41.84	13.17

Table 6: BLEU, ROUGE, and Cosine Score of facts retrieved from the OMCS Corpus against facts retrieved from the cumulative 20M corpus introduced by (Yu et al., 2022) using pre-trained ColBERTv2. Retrieved facts are compared to the ground truth explanation. Scores reported on 20K random samples from the test set of CRIC.

Figure 2: Examples illustrating pre-trained COLBERTv2 performs better than FAISS vector search in retrieving contextually relevant facts from OMCS Corpus. $C_{i,q}$ indicates i^{th} fact retrieved using question as the query by the pre-trained ColBERTv2, $F_{i,q}$ indicates i^{th} fact retrieved using question as the query by FAISS vector search.

the facts are about objects related to producing or listening to music. In (b), only the first fact retrieved by FAISS is close to what the query asks for, and all other facts are only connected to the query by one concept (such as 'riding' or 'walk'). The facts retrieved by ColBERT are directly related to sidewalks and activities it can be used for. Although none of them are direct answers, they are still semantically much closer to the query than the FAISS-retrieved facts. A similar pattern is noticed in (c) wherein the FAISS-retrieved facts are only related to the query by a singular concept (such as "branch" or "tree") but the ColBERT facts are both conceptually and contextually relevant; in this case, they are direct answers to the query.

E Additional Examples for Type-0, Type-1, Type-5 Explanations

Here we provide more examples of the different **Types** of explanation produced by Llama-3 index (a - g). Of the 8 examples presented, only in (c)

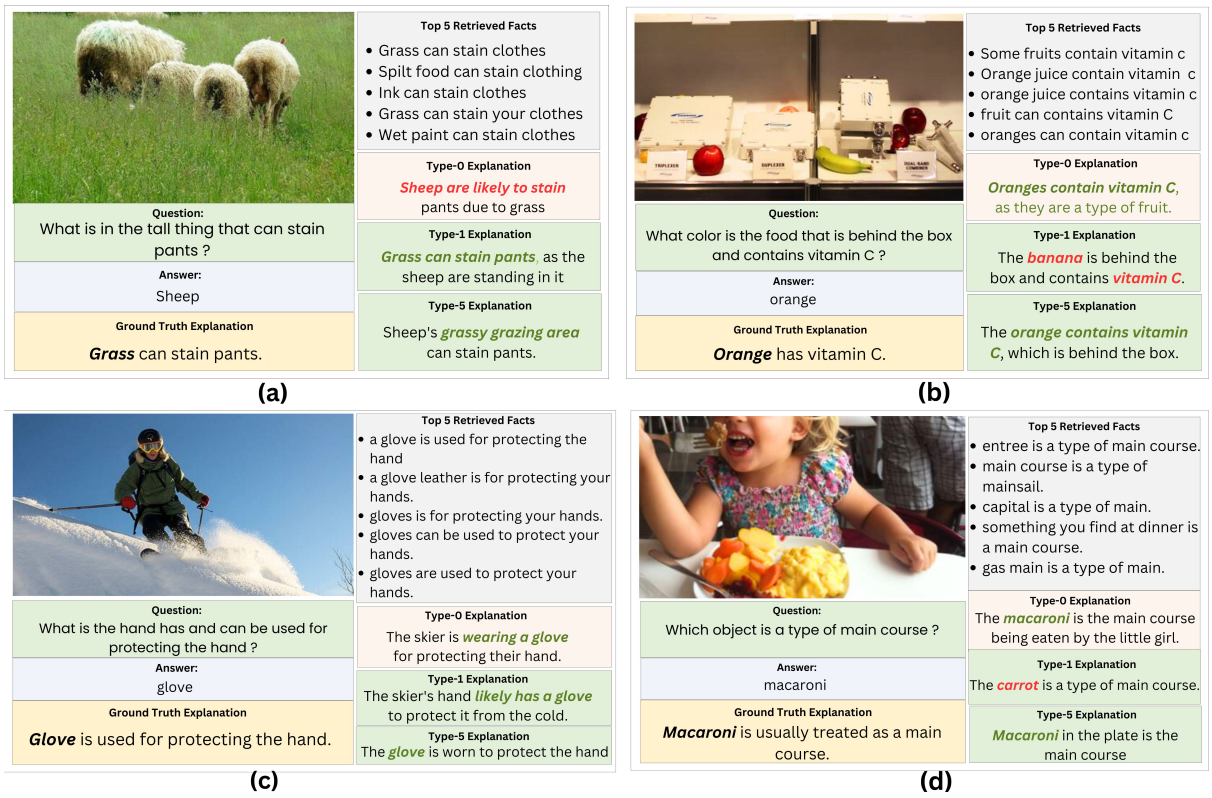

Figure 3: Examples from CRIC dataset showing the difference between the Ground Truth, Type-0, Type-1 and Type-5 Explanation generated by the LLama-3. In all four cases, the **Type-5** explanation is the most rational and contextually relevant. **Type-0** and **Type-1** explanations are often hallucinated (to align the explanation with the label) or do not make rational sense. More examples in the Appendix.

the **Type-5** explanation is incorrect. Even for that case, the explanation is still rationally correct because there is indeed no “object" on the airplane– the question refers to the text “FOX" on the engine which is technically not an object. In all the examples, the **Type-0** explanations are not semantically consistent; they stray off the context very quickly and produce factually incorrect claims. This is evidenced by the examples in Fig.7. The **Type-1** explanations, although consistent, are nearly never precise enough to enable the downstream VLM to identify the correct answer.

F Florence Finetuning For Explanation Generation And Comparison.

As we explore the generation capability of Florence (sVLM) with 606M parameters, we found that, though it excels in several vision-language task but it failed to generate a well reasoned explanation with respect to the context that is image and questions supplied to it. To ensure the quality of data to finetune Florence we used training split of the VQAe dataset comprising of 181K data instances. We used VQAe (Li et al., 2018) because of the fact it closely resembles the type of explanations that CRIC and eSNLI-VE has. We report the florence finetuned performance on CRIC in 2. Below we show the difference of some of the explanations generated by the florence vs LLAMA3 (**Type 5**).

G LLM Prompts for Explanation Generation: LLaMA3-8B & GPT-3.5

Similar to the Llama-3 prompt discussed earlier, a prompt was constructed for querying GPT3.5 to obtain refined explanations for our set of data. We feed GPT3.5 with Image captions, Objects Detected, Retrieved Facts from OMCS then ask it to analyze and generate an short explanation leveraging these details. We kept on improving our prompts through experimentation because, GPT/LLAMA generated extra texts containing additional reasoning for the explanation it generated, which are difficult to post-process. Below we illustrate the prompt template. To avoid label leakage in the generated explanations we also explicitly instruct the LLM to not produce the forbidden words. For instance terms like "image description", "caption"

Figure 4: More examples of **Type-0**, **Type-1**, and **Type-5** generated explanations.

confuse the VQA model, thus we filter out these words. This also includes words like "entailment", and "contradiction" when using the prompt to generate explanations for the e-SNLI-VE dataset. We

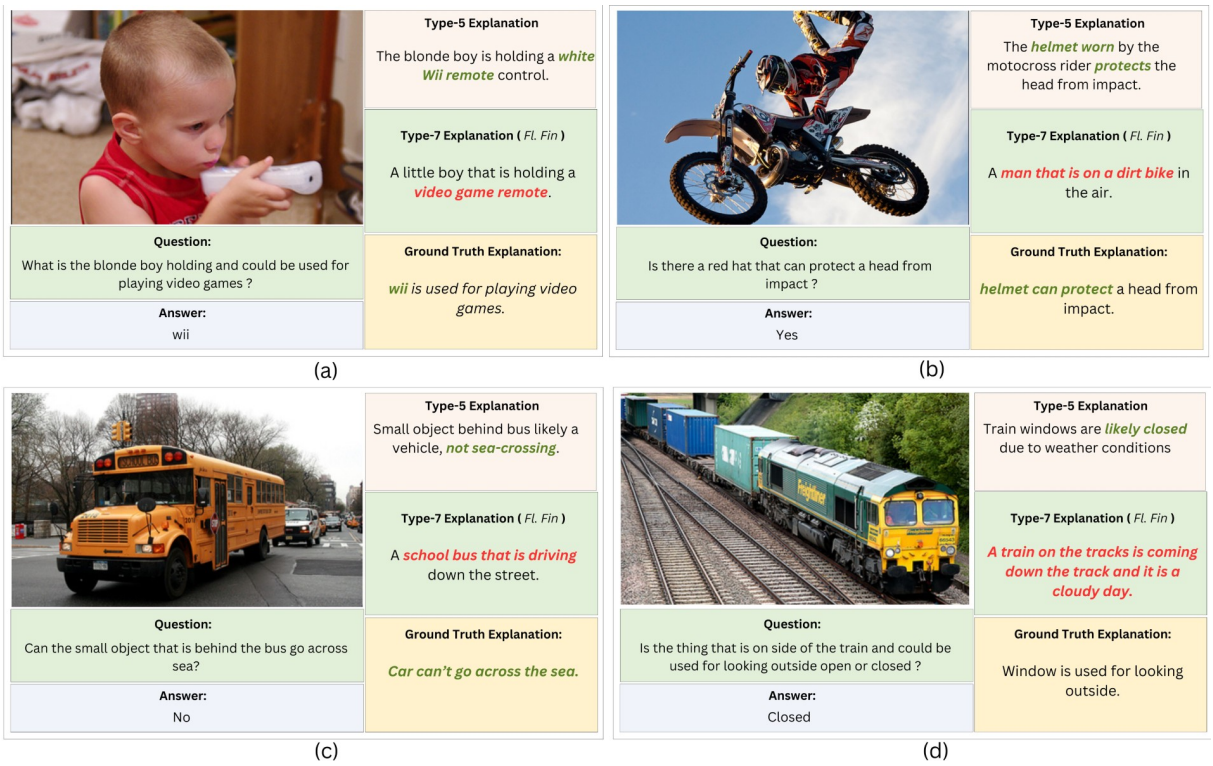

Figure 5: Examples showing difference between the LLAMA3 generated **Type 5** explanation and Florence Finetuned Generated explanations **Type 7**

Given an image description, you need to generate an explanation based on these informations.

Image Description: There are many different fruits and vegetables on the table
Objects Present: ['broccoli', 'apple', 'bowl']
Question: Is there a vegetable that is in the table and is large and contains vitamin c ?
Answer: No
Retrieved Facts are: ['Vegetable is related to vitamins.', 'Table is related to food prop.', 'Plant is related to vegetable description.', 'Vegetable is related to food.', 'Plant is related to vegetables.']

You strictly need to generate a short single line explanation within 10-15 words without any extra words, which can help VQA models derive the conclusion and nothing else. You might focus on the Retrieved Facts if information is not available in the objects detected and scene graphs.

##################
The forbidden words are: "image description", "captions". Dont start with "Here is single-line explanation", "Possible single-line explanation", "Based on the provided information, the explanation is:".

Figure 6: Prompt that we used to query GPT 3.5/LLAMA3 to obtain the all **Types** of explanation

explicitly instruct the LLM to generate only explanations within 10-15 words because in the knowledge concatenation experiments, we found that the longer the input text gets the lesser the performance the model delivers.

H Effect of Increasing Number of Facts

Retrieved facts, while somewhat useful, generally perform the worst in enhancing model performance, especially as more facts are concatenated. The primary hindrance is their often disjoint nature and lack of direct relevance to the specific question. In our experiments, as the number of concatenated facts increased, we consistently observed a decline in accuracy across various datasets. This decline can be attributed to several factors: the relevance of retrieved facts diminishes with quantity, introducing noise and reducing clarity. For instance, accuracy in the CRIC dataset dropped from 73.48% with just the question to 38.30% with five facts appended to the question. Similarly, ViLT's performance on e-SNLI-VE dataset fell from 76.46% to 69.62%. The inclusion of more facts tends to clutter the input with irrelevant or redundant information, making it difficult for the model to focus on the relevant information leading to degraded performance.

I Training Parameters and Resources

We report the learning rate and number of epochs below for the models we have used for VQA and Retrieval task from the knowledge bases. For all the experiments we have used dual NVIDIA-A40 GPU system each of 46 GB size.

Dataset	Knowledge Type	Split	Rougue-1	Rougue-2	Cumulative 1-gram	Cumulative 2-gram	Cumulative 3-gram	Cosine Score
e-SNLI-VE	Explanations	Train	24.35	5.97	13.20	6.31	3.88	**49.98**
	Retrieved Fact	Train	**27.11**	**8.06**	**20.92**	**10.58**	**6.62**	47.66
	Explanations	Val	25.79	6.43	14.50	6.85	4.20	**50.36**
	Retrieved Fact	Val	**31.87**	**7.81**	**20.17**	**10.04**	**6.26**	47.46
	Explanations	Test	25.95	6.54	14.50	6.85	4.17	**50.50**
	Retrieved Fact	Test	**32.08**	**7.98**	**20.30**	**10.14**	**6.34**	47.04
CRIC	Explanations	Train	31.76	13.78	7.16	4.86	3.70	54.40
	Retrieved Fact	Train	**80.34**	**69.99**	**75.52**	**69.57**	**63.16**	**80.33**
	Explanations	Val	29.73	12.49	6.27	4.15	3.14	51.59
	Retrieved Fact	Val	**80.85**	**70.34**	**75.99**	**69.95**	**63.52**	**81.37**
	Explanations	Test	29.77	12.45	6.32	4.19	3.17	51.64
	Retrieved Fact	Test	**80.95**	**70.51**	**76.06**	**70.05**	**63.64**	**81.49**

Table 7: LLM-generated explanations outperform the retrieved facts in assisting the models in deriving correct answers due to their relevant context, precision and coherence. However, the retrieved facts, while being less helpful for reasoning, yield higher BLEU & ROUGE scores than the generated explanations. This is attributed to the nature of BLEU and ROUGE metrics– retrieved facts have more overlapping n-grams with the ground-truth than the generated explanations even if they lack contextual relevance and depth.

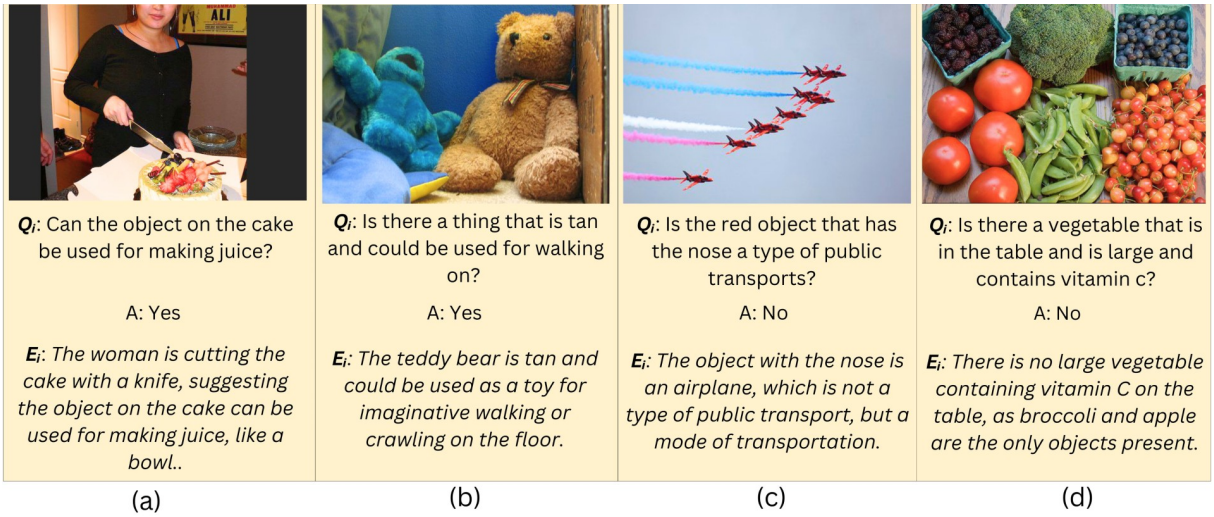

Figure 7: We report some instances where the LLM generates inaccurate **Type-0** explanations (where ground truth label was used). The visual and textual disconnects highlight the challenges LLMs face in reasoning, leading to 'hallucinations' that stray away from the visual context in the image..

Architecture	Input	CRIC	e-SNLI-VE
ViLT	Question	73.48	**76.46**
	1 Fact + Question	**73.44**	75.60
	2 Facts + Question	71.64	75.88
	3 Facts + Question	62.00	72.44
	5 Facts + Question	38.30	69.62
VisualBERT	Question	**69.53**	**74.48**
	1 Fact + Question	68.98	74.34
	2 Facts + Question	68.25	73.99
	3 Facts + Question	68.61	73.98
	5 Facts + Question	68.60	73.62

Table 8: Performance on varying the number of facts concatenated with the query.

Task	LR	Batch Size	#Epoch	#Hours
Fin. ViLT	5e-5	32	6	9.5
Fin. ViLT	4e-5	32	6	10
Fin. ColBERT	1e-4	32	4	2

Table 9: Hyperparameters for Finetuning tasks