NLKI: A *lightweight* Natural Language Knowledge Integration Framework for Improving Small VLMs in Commonsense VQA Tasks

Aritra Dutta IIT Kharagpur
Swapnanil Mukherjee Ashoka University
Deepanway Ghosal Independent Researcher *
Somak Aditya IIT Kharagpur

Abstract

Commonsense visual–question answering often hinges on knowledge that is missing from the image or the question. Small vision-language models (sVLMs) such as ViLT, VisualBERT and FLAVA therefore lag behind their larger generative counterparts. To study the effect of careful commonsense knowledge integration on sVLMs, we present an end-to-end framework (NLKI) that (*i*) retrieves natural language facts, (*ii*) prompts an LLM to craft natural language explanations, and (*iii*) feeds both signals to sVLMs respectively across two commonsense VQA datasets (CRIC, AOKVQA) and a visual-entailment dataset (e-SNLI-VE). Facts retrieved using a fine-tuned ColBERTv2 and an object information-enriched prompt yield explanations that largely reduce hallucinations, while increasing the end-to-end answer accuracy by up to 7% (across three datasets), making FLAVA and other models in NLKI match or exceed medium-sized VLMs, such as Qwen-2 VL-$2B$ and SmolVLM-$2.5B$. As these benchmarks contain 10–25% label noise, additional fine-tuning using noise-robust losses (such as symmetric cross-entropy and generalised cross-entropy) adds another 2.5% in CRIC and 5.5% in AOKVQA. Our findings expose when LLM-based commonsense knowledge beats retrieval from commonsense knowledge bases, how noise-aware training stabilises small models in the context of external knowledge augmentation, and why parameter-efficient commonsense reasoning is now within reach for $250M$ models. [1]

1 Introduction

Early Visual Question Answering (VQA) work already recognised that images and questions alone are often insufficient and called for the use of external knowledge sources (Wang et al., 2015; Aditya et al., 2018a, 2019). Factoid VQA now routinely benefits from graph- or web-scale retrieval (Wang et al., 2017; Marino et al., 2019; Chen et al., 2024), yet the commonsense variant remains under-explored, especially for *small* vision–language models (sVLMs) such as ViLT, VisualBERT and FLAVA. Recent attempts either bypass pre-trained small VLMs (Ye et al., 2023) or limit themselves to factual QA (Lin and Byrne, 2022; Yang et al., 2023). A big challenge in commonsense knowledge retrieval is a lack of a unified source – no single knowledge graph, corpus, or model covers everyday physics, social conventions, and object affordances. Large language models (LLMs) have emerged as promising but noisy reservoirs of such knowledge (Ghosal et al., 2023). These challenges motivate our four research questions: **RQ1)** *to what extent can commonsense*

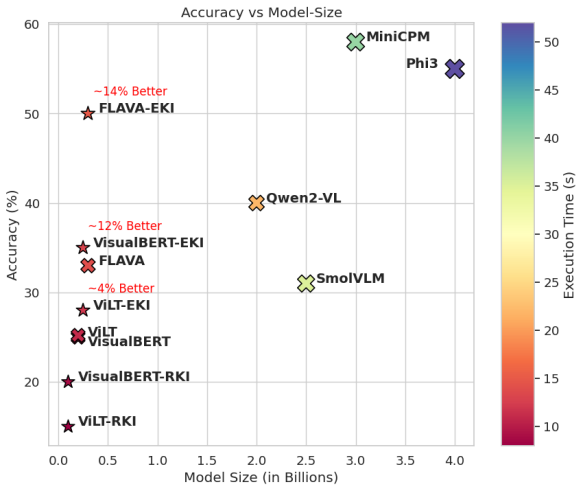

Figure 1: Comparison of accuracy and model size for various small Vision-Language Models (sVLMs) AOKVQA (Val). EKI: Type-5 Explanation Knowledge Integrated models, as defined in 2, while RKI: Retrieved Knowledge Integrated models. The colour gradient represents execution time, with red indicating longer durations and blue indicating shorter ones.

*Now at Google

[1]We release our code and checkpoints at: https://github.com/beingdutta/NLKI-Lightweight-Natural-Language-Knowledge-Integration-Framework

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 10543–10563
November 4-9, 2025

facts retrieved from knowledge bases and LLM-generated explanations help small VLMs?; And **RQ2)** *is one better suited than the other for commonsense reasoning in VQA systems?*; **RQ3)** *can noise-robust losses mitigate the pervasive label noise in the context of external knowledge augmentation?*; and to draw relevance to recent progress in lightweight generative AI, we analyse **RQ4)** *how do medium-size generative VLMs ($\leq 4B$) fare in commonsense tasks, in comparison to sVLMs (with or without knowledge integration)?*

To probe **RQ1–2**, we test the three sVLMs above ($\leq 240M$ parameters) on CRIC and AOKVQA (Gao et al., 2023; Schwenk et al., 2022) along with a visual entailment dataset, e-SNLI-VE (Xie et al., 2019). Using ground-truth explanations (along with the image and question) lifts accuracy by up to *10 to 12%*, signalling the necessity and scope for utilising text-based knowledge. We therefore contrast (*i*) ColBERTv2-retrieved facts from commonsense Knowledge Bases (KBs) and (*ii*) free-form textual explanation produced by Llama-3-8B from dense/region captions, list of objects, and retrieved facts. A richer visual context helps curb hallucination, while simple architectural tweaks temper the effect of occasional noisy snippets. Because label noise dominates these datasets, **RQ3** tests whether Generalised Cross-Entropy (GCE) (Zhang and Sabuncu, 2018) and Symmetric Cross-Entropy (SCE) (Wang et al., 2019) preserve the gains and further boost the performance of knowledge integration, while mitigating the effects of label noise. Finally, addressing **RQ4**, we evaluate Qwen-2, Phi-3-Vision, MiniCPM and SmolVLM ($\leq 4B$) and show that they too lack commonsense without explicit knowledge integration, underscoring the urgency of lightweight, knowledge-aware methods. Overall, we make the following contributions:

- **Plug-and-Play Knowledge Integration (NLKI)**: We introduce **N**atural-**L**anguage **K**nowledge **I**ntegration (NLKI) – a plug-and-play framework that combines a retriever, an LLM explainer, and a lightweight reader with any sub-$240M$ VLM, so each module can be analysed and improved independently to improve the overall accuracy of small VLMs in commonsense reasoning tasks.
- **Dense retrieval benchmark**: Compared to various state-of-the-art retrievers, we showed that finetuning ColBERTv2 using contrastive learning delivers the highest recall in retrieving the most relevant commonsense facts for CRIC, AOKVQA, and e-SNLI-VE queries.
- **Accurate LLM Explanation using Context and Knowledge-Enriched Prompting**: We show that adding object and region information of an image (through dense/region captions generated from Florence-2-large) along with the retrieved facts to the Llama-3.1-8B prompt sharply reduces noise and hallucinated details in the generated explanation, compared to a caption-only context.
- **Knowledge & noise-robustness gains**: LLM-generated explanations coupled with Symmetric Cross Entropy (SCE) or Generalised Cross-Entropy (GCE) preserve most of that gain under heavy label noise scenarios, raising AOKVQA accuracy by an average of 13.6% and CRIC, e-SNLI-VE by $2-4\%$ across three architectures;
- **Scale comparison**: Evaluations of Qwen-2, Phi-3-Vision, MiniCPM, and SmolVLM ($\leq 4B$) show these larger models still lack commonsense, while NLKI-equipped small VLMs ($\leq 240M$) can match or outperform them in some cases.

2 Related Work

Knowledge-based VQA systems pose a challenging task, as they need the models to comprehend visual and textual data while utilising external knowledge to derive correct answers. For instance, as depicted in Fig. 2, *"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., 2017; Zheng et al., 2021; Chen et al., 2021). Beyond KGs, logic-based formalisms have also been used to combine knowledge with reasoning, e.g., Probabilistic Soft Logic for visual puzzle solving in (Aditya et al., 2018b). For unstructured knowledge, Karpukhin et al. (2020) showed that the Dense-Passage-Retrieval (DPR) technique performs 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. Retrieval-augmented learning is widely used in language modelling and is increasingly explored in vision-language and multimodal tasks. Recent studies apply retrieval-based augmentation in several ways: (i) Image-based retrieval: Rao et al. (2024) retrieved image-caption pairs using a CLIP-based (Radford et al., 2021) scoring technique, processed by an OFA model ($182M$ parameters) for text generation. Wang et al. (2023) integrated similar images into vision-language attention layers in a transformer decoder. Rao et al. (2023) retrieved knowledge graphs (KGs), encoded them with a graph encoder, and used them in a BERT + ViT-B vision-language model. These approaches show promising results by enhancing VQA and image captioning through retrieval mechanisms.
(ii) Multimodal retrieval: Hu et al. (2023) and Yasunaga et al. (2023) proposed multimodal retrieval techniques compatible with language modeling, with Hu et al. (2023) using T5 + ViT ($400M$–$2.1B$ parameters). Caffagni et al. (2024) proposed a two-stage retrieval: CLIP retrieves documents, and Contriever extracts passages for LLaVA VLMs (7B+ parameters) in visual QA. While similar, our approach incorporates additional textual cues, expands retrieval beyond Wikipedia, and uses a smaller VLM.

3 Problem Formulation

In Visual Question Answering (VQA), some questions require knowledge beyond what is directly visible in an image. For example, identifying a giraffe as an *"even-toed ungulate"* or recognising a sofa as *"furniture used for sitting"* requires commonsense reasoning. Ambiguities in commonsense datasets further challenge models, as multiple correct answers may exist, making implicit knowledge crucial for accurate predictions. Addressing these gaps is essential for improving the VQA system's contextual understanding. 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 utilising a set of implicitly required commonsense knowledge facts (s) (K'). We utilise 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 pre-trained 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 a retriever, which is tuned to retrieve top k relevant facts ($\mathcal{F}_{ret}$), and $p_\Phi(\cdot)$ is termed as the reader, which utilizes the question, image, and retrieved facts to predict the answer. To generalise to LLM-generated explanations, we can think of the reader as a model that outputs a paragraph relevant to the question. We will utilise 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 Integrating Knowledge from Commonsense Corpus

Commonsense Corpus. Commonsense knowledge comprises basic facts and anecdotes about everyday objects, actions, and events, enabling humans to make inferences (Li et al., 2022). Since Transformer models reason well with text-based commonsense knowledge (Clark et al., 2021), we adopt Natural Language Knowledge Integration (NLKI), which is more flexible than structured knowledge graphs. Yu et al. (2022) introduced a 20M-fact natural language commonsense corpus from multiple human-annotated sources and web data. Based on initial ablations (Appendix Tables 7 and 8), we selected a subset: Open Mind Commonsense (OMCS) (Havasi et al., 2010) for its higher similarity scores with ground-truth explanations and its compact size (∼1.5M facts) describing everyday events and objects.

Commonsense Knowledge Retrieval. Given an image-question pair, the goal is to retrieve k relevant commonsense facts ($f_{i\leq k}$) from the commonsense corpus ($\mathcal{K}$) to fill in missing contextual knowledge. Ideally, retrieval involves abductive reasoning to infer necessary deductions. We primarily

explore text-based dense retrieval using SBERT (Reimers and Gurevych, 2019), ColBERTv2 (Khattab and Zaharia, 2020), and Stella selected from the MTEB leaderboard (Muennighoff et al., 2023).

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 embed the commonsense corpus ($\mathcal{K}$), indexing it with FAISS (Douze et al., 2024). Queries (e.g., question/hypothesis) are encoded using the same model, and Nearest Neighbour Search retrieves the top-k closest facts. Since SBERT captures semantic similarity, we further employ ColBERTv2, optimised for retrieval. Improved results (Table 9) led us to fine-tune ColBERTv2 for commonsense fact retrieval.

Figure 2: 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.

Finetuning ColBERTv2. As shown in Fig. 2, pre-trained ColBERTv2 retrieves partially relevant facts since it is trained on generalist datasets (MS-MARCO, TREC-CAR). To improve relevance, we fine-tune it following Khattab and Zaharia (2020), formatting the commonsense corpus into 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 (see Appendix C for finetuning details).

Commonsense Knowledge Integration. The retriever-reader architectures can be trained in two ways: (*i*) in an end-to-end fashion where the retriever and the VLM are finetuned together (Sachan et al., 2024), and (*ii*) where they are finetuned independently of each other. For simplicity, we go with the latter approach to improve both independently, as this allows for a modular plug-and-play framework wherein the retriever/reader can be switched as per the specific task. And in the case of sub-optimal retrieval performance, the reader ($p_\Phi(\cdot)$) is expected to learn to reason with noisy added knowledge.

4.2 Integrating LLM-generated Explanations

Using the image I, we extract various pieces of information such as multiple types of captions, a list of objects, and a scene graph to textually represent the visual context, which is also sometimes paired with the retrieved facts (these can be perceived as parameters to the prompt). To define an oracle setting (**Type 0**), we also supply the ground-truth label (GT-L). Interestingly, Llama-3.1-8B produces high-quality explanations when provided with the ground-truth label. Manual analysis shows that such explanations are riddled with hallucinations when the captions do not capture enough question-relevant information or the label is noisy in the case of Type-0 (See Fig. 10). Therefore, we performed a detailed analysis of the effect of varying all of the above parameters in generating an explanation. We provide the common instruction in Fig. I, as utilised for the instruction-finetuned LLMs (such as Llama-3.1-8B) to generate explanations[2]. The variants are the following:

- **Type-0**: TC + Q + $RF_{I,Q}$ + GT-L
- **Type-1**: TC + Q + $RF_{I,Q}$
- **Type-2**: DC + Q + $RF_{I,Q}$

[2]For Llama-3.1-8B we used the official implementation from Huggingface at https://huggingface.co/meta-llama/Llama-3.1-8B

- **Type-3**: RC + Q + $RF_{I,Q}$
- **Type-4**: RC + Q + O + $RF_{I,Q}$
- **Type-5**: DC + RC + Q + O + $RF_{I,Q}$
- **Type-6**: DC + RC + Q + O (Type 5 without knowledge facts).
- **Type-7**: Florence Finetuned Explanations,

where TC: Traditional Image Caption, DC: Dense Caption, RC: Region Based Caption, Q: Question, RF: Retrieved Facts, O: Objects, GT-L: Ground Truth Label.

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

In this work, we choose two popular VQA and one Natural Language Inference (NLI) datasets that require commonsense reasoning: CRIC (Gao et al., 2023), AOKVQA (Schwenk et al., 2022) and e-SNLI-VE (Xie et al., 2019) due to their size, diversity of questions, and availability of ground truth explanations. In this work, we do not consider datasets that test (encyclopedic) knowledge about entities such as FVQA (Wang et al., 2017), KB-VQA (Wang et al., 2015). More details about our choice of datasets are in Appendix A.

5.2 Architectures & Training

We choose three small-vision-Language models – VisualBERT (Li et al., 2019), ViLT (Kim et al., 2021), FLAVA (Singh et al., 2022) – spanning two broad modality fusion architectures, all under *240M* parameter size, including one dual stream architecture, and two single stream architectures. We briefly describe each of these architectures for our VQA task in Appendix B. For finetuning, we use the standard cross-entropy loss objective (and extend this with noise-robust variants in §6.5). We have tried multiple variants to integrate commonsense knowledge, such as *majority voting* (details in Appendix C), *string concatenation*, and their variants with noise-robust loss functions.

6 Results

6.1 Analysis of Standalone Retrieval Performance

We evaluate retrieval methods to identify the best retriever component of our *NLKI pipeline*. We compare the relevance of retrieved facts (by each of the retrievers) with ground-truth explanations for each of the datasets. As per our studies specific to commonsense knowledge, FAISS, which relies on pre-trained SBERT embeddings, struggles with retrieving relevant *missing* information. Fig. 2 shows how the facts retrieved by FAISS are conceptually close to the question but not relevant, whereas the facts retrieved by ColBERTv2 are exact. In contrast, ColBERTv2 retrieves more precise facts and significantly outperforms FAISS and other state-of-the-art retrievers like Stella-400M[3] across BLEU, ROUGE, and cosine similarity metrics (see Tab. 9).

6.2 Capturing Visual Context for LLM-Generated Explanations

Table 1 compares prompt settings for explanation generation. Type-5, incorporating dense captions, region captions, retrieved facts, and objects, outperforms other variants, effectively capturing colour information and context. Our manual analysis of baselines suggested that sVLMs struggle with noisy, ambiguous and colour-based questions. As the region and dense captions capture colour information, Type-5 explanations prove to be the most effective and relevant (compared to generated explanations of other *Types* and retrieved facts standalone). We further explored Smaller LLaMA 3.2 models (3B & 1B) to generate explanations. Such explanations turned out to be noisier and less accurate with poorer visual context, making them unsuitable (Detailed results in Appendix H).

6.3 Baseline Performance

To estimate the vanilla performance (without knowledge augmentation), we finetuned ViLT, VisualBERT, and FLAVA with the image, the question (hypothesis) and the answer (label) across all our datasets. We report the results in Table 2. Manual analysis (see §6.5) reveals that the sub-optimal result is due to 1) *lack of day-to-day commonsense knowledge* and 2) *noise and ambiguity* in the datasets. We found that CRIC has a non-negligible amount of label noise (see Figs. 6, 4) across all the splits, which caused the models to learn erroneous patterns in the data. AOKVQA is comparatively less noisy but follows a similar noise distribution to CRIC (see Fig. 5). This necessitates the need for external commonsense knowledge, while build-

[3]We used the official HuggingFace implementation at https://huggingface.co/NovaSearch/stella_en_400M_v5

Type	Explanation	B-1	B-2	B-3	R-1	R-2	R-L	Cosine
Type-0	TC + Q + $F_{I,Q}$ + $y_{I,Q}$	18.60	11.41	8.32	36.01	12.88	32.08	53.21
Type-1	C + Q + $F_{I,Q}$	24.65	16.64	12.81	40.00	17.75	36.80	54.75
Type-2	DC + Q + $F_{I,Q}$	27.02	18.85	14.84	42.25	20.26	38.86	58.15
Type-3	RC + Q + $F_{I,Q}$	30.84	22.57	18.16	45.90	24.51	42.80	59.60
Type-4	RC + O + Q + $F_{I,Q}$	30.46	22.30	18.01	46.01	24.35	42.44	60.08
Type-5	DC + RC + O + Q + $F_{I,Q}$	**31.08**	**22.77**	**18.41**	**46.30**	**24.63**	**42.78**	**60.17**
Type-6	DC + RC + O + Q	25.88	18.51	14.67	41.93	20.86	38.57	57.01
Type-7	I + $< TP >$	7.20	2.97	1.87	11.76	00.75	10.49	22.39

Table 1: Results from CRIC (AOKVQA in App. H) comparing 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-k (**B-**k, $k \in \{1, 2, 3\}$), ROUGE-k (**R-**k, $k \in \{1, 2, L\}$), and Cosine similarity scores. Q: Question, O: Objects, TC: Traditional Image Caption, $F_{I,Q}$: Retrieved Facts, DC: Dense Caption, RC: Region Caption, I: Image and TP: Task Prompt. These scores can be compared with the retrieved facts score metrics **B-1**@5, **R-L**@5, **Cosine**@5 of Tab. 9

ing noise robustness to specialise the models in the commonsense VQA task.

6.4 End-to-End NLKI Performance

Table 2 shows that the full pipeline with **Type-5** explanations (with DC, RC, TC, O, RFs as the parameters of the prompt) is the strongest variant. This echoes the scores in Table 1, as **Type-5** rationales most closely match groundtruth explanation in both n-gram overlap and the semantic space than **Types 1–4**. The payoff is large: *+13%* for FLAVA on AOKVQA and *+2%* even on the noisy CRIC split. We also compared our *NLKI pipeline* with a specialised retrieval-augmented baseline, termed KAT (Gui et al., 2022). The details of the knowledge concatenation and truncation strategy in our pipeline are present in Appendix B. We trained the Knowledge-Augmented Transformer (KAT) with explicit and implicit knowledge; KAT[4] lags behind the NLKI **Type-5** variant on every dataset except ViLT with **Type-5**.

Architecture	Retrieved/ Generated	Accuracy		
		e-SNLI-VE	CRIC	AOKVQA
KAT	Expl. Knowledge	73.22	67.38	32.72
KAT	Expl. + Impl. Knowledge	**76.30**	**69.71**	**38.60**
ViLT	✗	76.46	72.99	24.01
Concat (ViLT)	**Type 5**	78.46	74.95	28.15
Concat (ViLT)	Type 7	64.66	63.34	20.08
Majority Voting (ViLT)	*Q/CB-FT*	74.51	69.08	13.45
Concat (ViLT)+*CE	**Type 5**	**78.57**	**76.98**	**33.45**
VisualBERT	✗	74.48	62.60	23.6
Concat (VisualBERT)	**Type 5**	78.83	64.69	35.40
Concat (VisualBERT)	Type 7	62.46	55.22	20.00
Majority Voting (VisualBERT)	*Q/CB-FT*	72.53	60.25	19.00
Concat(VB)+*CE	**Type 5**	**78.95**	**67.15**	**40.12**
FLAVA	✗	79.93	73.11	33.07
Concat (FLAVA)	**Type 5**	81.54	75.02	47.85
Concat (FLAVA)	Type 7	70.18	64.30	32.09
Majority Voting (FLAVA)	*Q/CB-FT*	72.53	60.25	32.68
Concat(FLAVA)+*CE	**Type 5**	**82.05**	**77.85**	**47.85**
	Gold-label baseline			
ViLT	Ground Truth Expl.	92.47	82.98	56.31
ViLT	Type-0	**97.91**	**97.02**	**58.38**
ViLT	Type 1	75.92	55.65	20.28
VisualBERT	Ground Truth Expl.	93.49	77.00	69.35
VisualBERT	Type-0	**97.89**	**96.60**	**87.16**
VisualBERT	Type 1	76.11	53.82	35.19
FLAVA	Ground Truth Expl.	93.04	82.36	72.65
FLAVA	Type-0	**98.11**	**97.34**	**90.03**
FLAVA	Type 1	69.29	50.45	39.71

Table 2: Performance of Baseline vs NLKI (ours) vs Gold-label baseline. Expl. and Impl. refer to *Explicit* and *Implicit* respectively. *Q/CB-FT* are facts retrieved by fine-tuned ColBERTv2 with the question as the query, **CE* rows contain the best performance when various noise robust loss function is applied with the **Type-5** explanation integration (see §6.5).

Effect of Noise. Further analysis shows that NLKI still inherits a lot of label noise. We manually analysed 1000 CRIC training samples (Fig. 4). This revealed five noise classes: label, image, question, image–question mismatch, and ambiguous labels (§G)—with label noise being the most frequent (180 among 1000). For example, the aircraft image in Fig. 6c is labelled "Grass" instead of "Water", so even a correct **Type-5** explanation fails; Fig. 6d allows two valid answers, blurring the target. Noise compounds when hallucinated explanations reinforce wrong labels (Fig.10d). We also

[4]Implementation details in Appendix B.

clearly see that the datasets have a varying degree of noise. On a similar analysis, nearly 90 instances (out of 1000) in AOKVQA answers were unanimously judged ambiguous or incorrect by three independent raters.[5] In contrast, e-SNLI-VE natural language inference benchmark contains three NLI answer labels, considerably less noisy than VQA datasets. Hence, we next incorporate noise-robust loss functions to prevent small 240M VLMs from over-fitting to noisy outliers in §6.5. With noise-robust loss functions, all architectures surpass KAT's performance by a large margin (except ViLT on AOKVQA), underscoring the competitiveness of our lightweight scheme.

6.5 Building Noise Robustness

We replace the default cross-entropy (CE) objective with *down-weight* or *smoothen* the contribution of mislabeled samples. For each example image and question (x_i), target answer (y_i), we use the standard cross-entropy loss objective to fine-tune each model. Assume, for i^{th} example, the model logits are $z_i \in \mathbb{R}^C$ (C=#classes), $p_i = \mathrm{softmax}(z_i)$ and $p_y^{(i)}$ denote the predicted probability of the correct class. The standard loss becomes $\mathcal{L}_{\mathrm{CE}} = -\sum_i \log p_y^{(i)}$. The noise-robust variants can be defined as follows.

Symmetric Cross-Entropy (SCE). Following Wang et al. (2019), we combine the usual cross-entropy with a *reverse* term that penalizes over-confident errors:

$$\mathcal{L}_{\mathrm{RCE}}(p,y) = -\sum_{c=1}^{C} p_c \log y_c = -A\sum_{c\neq y} p_c \quad (1)$$

$$= -A\,(1-p_y), \quad (2)$$

$$\mathcal{L}_{\mathrm{SCE}}(p,y) = \alpha\,\mathcal{L}_{\mathrm{CE}}(p,y) + \beta\,\mathcal{L}_{\mathrm{RCE}}(p,y), \quad (3)$$

where p_y is the predicted probability of the ground-truth class, we fix $\log 0 = -\gamma$ with $\gamma = 4$, $\alpha = 0.1$, $\beta = 1.0$ (the values of α and β are directly taken from (Wang et al., 2019)). The CE component ensures fast convergence, while the $-\gamma(1-p_y)$ term flattens the loss landscape around noisy labels, granting robustness to the 10 % to 30 % label noise observed in CRIC and AOKVQA.

Generalised Cross-Entropy (GCE). Zhang and Sabuncu (2018) proposed

$$\mathcal{L}_{\mathrm{GCE}}(p,y) = \frac{1-p_y^q}{q}, \quad 0 < q \leq 1,$$

which reduces to mean absolute error (MAE) as $q \to 0$ and to standard CE as $q \to 1$. We use $q = 0.7$ directly from (Zhang and Sabuncu, 2018), and form a convex mixture with CE:

$$\mathcal{L}_{\mathrm{Mixed}}(p,y) = \lambda\,\mathcal{L}_{\mathrm{CE}}(p,y) + (1-\lambda)\,\mathcal{L}_{\mathrm{GCE}}(p,y)$$

We use $\lambda = 0.4$ for noisy datasets (with 0.9 for e-SNLI-VE, which has limited label noise).

Training protocol. For each loss function, we train with vanilla CE for two epochs to stabilise early gradients, then switch to the chosen robust loss. We keep the batch size, optimiser, and learning-rate schedule identical to the CE baseline to isolate the effect of the loss.

6.6 Effect of Noise-robust Training

Table 3 compares CE, SCE, and CE + GCE on all three datasets for ViLT, VisualBERT and FLAVA for CRIC. On **CRIC**, **SCE** lifts FLAVA by **+2.8%**, VisualBERT by **+2.46%** and ViLT by **+2.03%** over CE, confirming that SCE's reverse term curtails over-fitting to noisy labels. On **AOKVQA**, **SCE** delivers the best trade-off (**+5.5%**), having low to moderate noise with ViLT and VisualBERT (Table 12 in the appendix). But FLAVA already reaches high performance with vanilla CE, so the extra regularisation offered by SCE or GCE brings only a marginal change (-0.12 %) and can even over-smooth (GCE+CE). On **e-SNLI-VE** standard CE *outperforms* its robust siblings. Raising q ($0.7 \to 0.95$) or λ ($0.4 \to 0.9$) makes CE + GCE converge to CE and closes the gap (Table 11 in the appendix).

Architecture	Type	CE	SCE	GCE+CE
ViLT	Type-5	74.95	**76.98**	75.13
VisualBERT	Type-5	64.69	**67.15**	65.66
FLAVA	Type-5	75.02	**77.85**	76.98

Table 3: Comparison of performance (in percentage) for different architectures using various loss functions on the CRIC test split of 76K samples. For e-SNLI-VE and AOKVQA, refer to Tables 11 and 12 in the Appendix.

In our noise ablation (Figs. 4 & 5, Appendix G), SCE delivered the largest gains in conditions of high label noises, mirroring the threshold reported in Wang et al. (2019)—whereas a 0.4 CE + 0.6 GCE mix performed the best in moderate level noise conditions, and standard CE remained optimal on e-SNLI-VE with only 3 labels. These

[5] Details in Appendix G.

findings explain why prior work that applied a single loss across heterogeneous benchmarks reported inconsistent gains; by adapting the loss to the dataset (according to the level of label-noise), we achieve stable improvements without any architectural change or extra inference cost.

7 Performance of Generative VLMs

Models	EM	P	R	F1	ACC
Qwen2-VL	31.18	40.00	92.82	93.17	41.90
SmolVLM	10.48	31.42	89.91	90.54	33.89
MiniCPM	**58.60**	**96.05**	**95.7**	**95.83**	**58.58**
Phi3-Vision	52.64	96.06	95.28	95.61	53.24

Table 4: Performance of Generative Models on AOKVQA val split where *EM* is exact string match *P* is Precision, *R* is Recall, *F1* is the F1 score, and *ACC* is the cosine similarity score.

We further benchmark (smaller, less than $4B$) generative instruction-tuned models such as Qwen2-VL ($2B$) (Wang et al., 2024), Phi3-Vision ($4.1B$) (Abdin et al., 2024), MiniCPM ($3.43B$) (Hu et al., 2024) and SmolVLM ($2.25B$) on AOKVQA. From the results in Table 4, we observe that **NLKI + SCE** turns a $240M$-param FLAVA into a model that beats 2B-param Qwen-VL and SmolVLM on AOKVQA with far less compute time and power.

8 Discussion

Learnings from Gold-label baseline Setup and Effect of Additional Context on Accuracy. Table 2 confirms that label-supervised **Type-0** explanations push accuracy far beyond both ground-truth texts and **Type-1** (no-label) variants. Injecting the answer often *repairs* missing context—see Fig.7b—and yields more assertive wording (e.g., "*the macaroni is the main course*" in Fig.7d). Yet the benefit is fragile: in a 1.5K sample drawn across all datasets, 51% of **Type-0** explanations still hallucinate or add spurious detail because they rely on weak visual cues. By contrast, our gold-label-free **Type-5** explanation prompt, which feeds dense and region captions plus retrieved facts, cuts that rate to 18.5% and grounds over 80% of explanations in the scene (Fig. 10; definition in Appendix H). A Florence-finetuned captioner used in place of the LLM underperforms; its explanations are shorter, omit causal links, and have lower end accuracy (Table 1; examples in Fig. 9).

When does knowledge integration help? Table 2 demonstrates that *NLKI alone without noise mitigation*, particularly with **Type-5** explanations, has a marginally positive impact on model performance for CRIC, while having a moderate to significant impact in e-SNLI-VE & AOKVQA. Some key observations emerge from our analysis of the NLKI framework: **1.** *The better the generated explanations are, the better the effectiveness of NLKI, which even outperforms the complex integration pipelines like KAT*. **2.** *To frame a quality Natural Language knowledge that can be effectively utilised, the explanations should capture relevant visual context in textual form while minimising hallucinations*. **3.** *NLKI, when coupled with noise robust losses, matches and even outperforms some of the performance of the architectures in the range of 1-4 B parameters*. **4.** *The cleaner the dataset and the stronger the model is, the smaller the benefit of noise-robust losses due to less chance of overfitting on mislabeled samples*. For example, on AOKVQA, they mainly rescue the lighter architectures while leaving FLAVA nearly unchanged. **5.** *If ground-truth (often noisy) labels are used to generate an explanation (gold-label baseline), they are more prone to hallucination, leading to unreliable predictions, as in Fig. 10*. If raw knowledge retrieved from knowledge bases or graphs is used as guidance, it may introduce additional noise and hinder performance rather than improving it. Table 10 shows our ablation study on augmenting k facts to questions degrades performance as k increases.

Impact in the context of Generative VLMs. Table 4 shows that despite rigorous pretraining and the use of advanced architectures, generative models still struggle with visual commonsense reasoning tasks. The best-performing model, MiniCPM, achieves only 58.60%, followed by Phi-3 Vision. This highlights the persistent lack of commonsense knowledge in medium-sized VLMs, as they fail to outperform the NLKI coupled with the noise-robust loss framework when applied to encoder-only models such as FLAVA, ViLT, and VisualBERT. We plan to explore the impact of NLKI in generative models as part of future work.

Inference Time Latency of NLKI Framework. While we highlight NLKI's empirical gains, it is also important to consider its computational footprint. Our framework introduces additional modules beyond the base VLM, namely, captioners, an object detector, a retriever and an explainer—so we

report the full breakdown of latency, FLOPs, and GPU usage in the table below. The total pipeline latency for a single image–question pair is 1.32s when components are run sequentially, with most of the cost concentrated in captioning and explanation generation. The reader stage itself is negligible ($\leq$ 65 ms across all tested VLMs). Running captioning and object detection concurrently shortens the critical path to 0.87s ($\approx$ 34% faster), though this raises peak memory load from $\approx$ 5 GB to $\approx$ 15 GB. Importantly, the retriever and captioners can be executed offline or on CPU, and the explainer can be swapped for a smaller LLM (e.g., Llama 1B or 3B), further reducing the online GPU footprint. Thus, although NLKI involves multiple stages, its design remains lightweight and deployable, offering a favourable trade-off between efficiency and the substantial performance gains reported in 6.

Tasks	Latency	FLOPS	GPU Usage
Dense Cap.	235.80	1680	5.2
Region Cap.	313.75	1680	5.2
Object Det.	225.06	1680	5.2
Expl. Gen.	486.53	735.48	15
KB Retrieval	114.12	7	0.78
Answer Prediction	65.02	55.84	1.9
Overall Pipeline Latency: (235.80 + 313.75 + 225.06 + 486.53 + 65.02) = 1.32 sec			

Table 5: GPU usage (in GB), wall-clock latency (in ms), and theoretical FLOPs (in GFLOPS) for the NLKI framework. *Cap.* stands for Captioning, *Det.* stands for Detection, *Expl. Gen.* stands for Explanation Generation.

9 Acknowledgements

We gratefully acknowledge the full support of the Science and Engineering Research Board (SERB), Government of India, through the Start-up Research Grant SRG/2022/000648 (PI: Somak Aditya) for this work.

10 Conclusion

Our study shows that pairing external commonsense knowledge with noise-aware training lifts $\leq$240M vision-language models to the tier of far larger systems. A lightweight NLKI pipeline– ColBERT-v2 retrieval plus Llama-3.1-8B generated **Type-5** explanations– boosts the performance of ViLT from 24 $\rightarrow$ 38 and FLAVA from 46 $\rightarrow$ 54 % on AOKVQA, outmatching 1–4B parameter generative baselines at a fraction of the compute. Noise audits reveal 17–23% faulty labels in CRIC and AOKVQA (virtually none in e-SNLI-VE); swapping vanilla CE loss for Symmetric CE adds up to +3%, while a CE + GCE mix is best at moderate noise, and vanilla CE suffices for clean data. Finetuned ColBERT raises retrieval precision, and enriching prompts with dense/region captions curbs hallucination, though one fact per question is optimal given VLM context limits. In sum, NLKI plus a dataset-aware robust loss turns 250M-parameter sVLMs into noise-resilient commonsense VQA engines that rival or beat multi-billion-parameter models without extra inference cost.

Limitations

Our pipeline is modular, but still serial retriever, explainer and reader are tuned in isolation; we do not explore joint optimisation. Noise robustness is tackled only at the label level with loss re-weighting; we leave open richer defences against other noises (such as in questions, image-question mismatch, etc). Our explanation module relies on Llama-3.1-8B, a single LLM family; whether larger or instruction-specialised models change the trade-offs is unknown. NLKI assumes access to accurate object lists and dense/region captions from auxiliary detectors, but the impact of errors in these vision tools is not assessed directly.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Somak Aditya, Yezhou Yang, and Chitta Baral. 2018a. Explicit reasoning over end-to-end neural architectures for visual question answering. *Preprint*, arXiv:1803.08896.

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.

Somak Aditya, Yezhou Yang, Chitta Baral, and Yiannis Aloimonos. 2018b. Combining knowledge and reasoning through probabilistic soft logic for image puzzle solving. In *Uncertainty in 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*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *Preprint*, arXiv:2401.08281.

Difei Gao, Ruiping Wang, Shiguang Shan, and Xilin Chen. 2023. Cric: A vqa dataset for compositional reasoning on vision and commonsense. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5561–5578.

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.

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.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *Preprint*, arXiv:2404.06395.

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.

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)*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. *Preprint*, arXiv:2210.07316.

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.

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.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. *Preprint*, arXiv:2206.01718.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. *Preprint*, arXiv:2112.04482.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

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, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2017. Fvqa: Fact-based visual question answering. *Preprint*, arXiv:1606.05433.

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*.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. 2019. Symmetric cross entropy for robust learning with noisy labels. *CoRR*, abs/1908.06112.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.

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 Datasets

We focus on CRIC, AOKVQA, and e-SNLI-VE, which emphasise open-ended, contextual commonsense reasoning over encyclopedic or synthetic factoid knowledge. Datasets like FVQA and WHOOPS, which centre on factual recall or synthetic errors, are excluded as they diverge from our goal of modelling everyday reasoning.

CRIC contains 494K automatically generated questions over 96K Visual Genome images, enriched with 3.4K knowledge items from ConceptNet and Wikipedia. **AOKVQA** includes 25K diverse, crowdsourced questions based on COCO images. Both CRIC and AOKVQA are multiple-choice tests with four options each. **e-SNLI-VE** combines Flickr30k and SNLI-VE, with hypotheses labeled as *Entailment*, *Contradiction*, or *Neutral*.

Dataset	Type	Train	Val	Test	Answer Type
CRIC	VQA	364K	76K	84K	MCQ
AOKVQA	VQA	17K	1.1K	6.7K	MCQ
e-SNLIVE	NLI	401K	14K	14K	NLI Labels

Table 6: Table showing the split size and answer type for all datasets.

B Architecture Description

We have selected models all under 240M for our VQA classification task. We opted for two different architectures 1. Single Stream 2. Dual Stream is based on the processing of both modalities, vision and text.

ViLT (Kim et al., 2021) 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.

VisualBERT (Li et al., 2019) requires a set of visual embeddings (representing the visual features) extracted from an image encoder and text embeddings encoded with BERT. Then, it utilises the self-attention mechanism for implicit alignment of elements in the input text and regions within the input image.

FLAVA (Singh et al., 2022) is a multimodal transformer-based model designed to process and align visual and textual information for various vision and language tasks.

KAT Implementation KAT originally used filtered English Wikidata and GPT-3 to extract implicit knowledge. In our setup, we use `gpt-3.5-turbo` for this step. Since VinVL and Oscar (used by KAT for extracting image features and captions) are no longer publicly available, we replace them with Florence-generated dense and region captions to query GPT-3.5. We retain the original prompt format but, due to resource constraints, combine answer and explanation generation into a single API call with a modified prompt. Final results are reported in Table 2.

Model Context Length & Truncation Strategy To integrate external knowledge into the model input, we prepend the generated explanation to the original question using the format: `<explanation>[SEP]<question>` before pairing this text with the corresponding image features. For the smaller reader VLMs, we enforce a strict 100-token limit on the concatenated explanation–question string: if the combined text exceeds this budget, truncation is applied from the end, ensuring that the leading tokens, which typically contain the most informative content, are preserved. This strategy was consistently applied across all datasets and models. Empirically, truncation was seldom triggered, as the average token counts remained well below the limit (e.g., CRIC: 12.74 tokens for questions and 19.35 for explanations; AOKVQA: 8.69 and 9.09; e-SNLI-VE: 7.39 and 10.43, respectively).

C Retrieval

What is majority voting? In the majority voting setup, we concatenate each of the top-5 retrieved facts individually with the question and image context to form five separate inputs. The model then predicts an answer for each of these fact-augmented inputs independently. We collect all five predicted labels and select the final answer by majority vote—i.e., the label that appears most frequently across the five runs. This approach helps smooth over noisy or irrelevant facts by relying on the consensus across multiple evidence sources, making the final prediction more robust to individual retrieval errors.

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 the CRIC and

e-SNLI-VE datasets. Although the CRIC test set is larger, a subset of 20K facts was chosen to make the results comparable to those 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 20 times smaller; hence, retrievals from the corpus are significantly faster.

Retrieval Performance for Commonsense VQA Tasks Here we present the detailed scores across the metrics like BLEU, ROUGE, and cosine similarity for top-5 and top-10 results for measuring the retrieval performance specific to our task. In Table. 9 we present the result of the three top retrievers for clarity, which was found to be effective for our commonsense VQA task.

Examples of Retrieved Facts: FAISS vs ColBERTv2 We provide more examples in Fig. 3, 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 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 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.

Finetuning Colbert Using a contrastive learning approach, we optimise the pairwise softmax cross-entropy loss over the relevance score $S_{q,d}$ 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, with a single 46 GB NVIDIA A40 GPU, having a maximum length of the facts as 128 (most facts are much smaller in length), and default settings for other parameters as in Khattab and Zaharia (2020). We used $160K$ samples from the training splits of CRICs for finetuning due to the variety of queries in it. We experimented with adding more than $160K$ data to finetune it, but it didn't help improve the scores any further.

D Effect of Increasing Number of Retrieved Facts

While somewhat useful, retrieved facts 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 72.99% with just the question to 38.30% with five facts appended to the question. Similarly, ViLT's performance on the e-SNLI-VE dataset fell from 76.46% to 69.62%. The same pattern repeats for other models as well. 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.

E Generative Small VLMs

As the research community increasingly prioritises efficiency and low carbon footprints, there has been a growing shift towards small-vision language models (sVLMs) and low-resource training strategies. Therefore, in the context of commonsense VQA, we benchmarked the mentioned generative instruction-tuned VLMs[6] on our datasets. We limit ourselves to models with < 4 billion parameters, as shown in Figure 1. The effectiveness of these rigorously pre-trained generative models in commonsense VQA remained underexplored. The observations in Table 4 suggest the reason for the sub-optimal *EM* and *Cosine Accuracy* scores is that

[6]We used the standard HuggingFace implementations for all the small generative VLMs: https://huggingface.co/Qwen/Qwen2-VL-2B-Instruct, https://huggingface.co/openbmb/MiniCPM-V, https://huggingface.co/HuggingFaceTB/SmolVLM-Instruct, https://huggingface.co/microsoft/Phi-3-vision-128k-instruct

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 7: 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 were 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 8: 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 were reported on 20K random samples from the test set of CRIC.

Query	Method	R-L@5	R-L@10	B-1@5	Cosine@5	Cosine@10
Q	SBERT + FAISS	**41.86**	**47.80**	**38.30**	**53.56**	**58.46**
C + Q	SBERT + FAISS	38.57	44.70	35.21	51.54	56.88
Objects + Q	SBERT + FAISS	40.93	46.73	37.34	52.61	57.67
SG + Q	SBERT + FAISS	36.05	41.91	32.88	47.01	52.21
All + Q	SBERT + 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**
C + Q	ColBERTv2-FT	52.03	55.77	45.95	64.34	68.39
Objects + Q	ColBERTv2-FT	52.11	56.26	46.27	66.01	69.41
SG + Q	ColBERTv2-FT	31.12	34.89	27.56	41.99	43.75
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
C + Q	Stella-en-v5	32.28	38.76	28.56	50.14	55.10
Objects + Q	Stella-en-v5	43.89	51.36	39.27	**62.98**	**67.66**
SG + Q	Stella-en-v5	30.02	36.50	27.50	50.02	53.65
All + Q	Stella-en-v5	29.01	34.65	25.67	46.76	51.15

Table 9: 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. Finetuned ColBERTv2 outperforms all other methods. Q: Question, C: Caption, O: Detected Objects

generative models can provide expressive answers based on their pre-training, but fail when asked questions that require explicit commonsense reasoning abilities. Also, it raises concerns about their ability to generalise beyond their training data, as in our case, no external knowledge was provided during this benchmarking. We used bert-score (Zhang et al., 2020) to evaluate the *P*, *R*, *F1*, which uses the contextual embeddings for computing token similarity. We chose the cosine similarity threshold as *0.71* based on manual analysis of results specific to our datasets.

F Qualitative Analysis of Generated, Retrieved and Manual Explanations.

In Fig. 7, we show ground truth, retrieved facts, and generated explanations (**Types 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 cap-

Figure 3: Examples illustrating that 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 the question as the query by the pre-trained ColBERTv2, $F_{i,q}$ indicates i^{th} fact retrieved using a question as the query by FAISS vector search.

Architecture	Input	e-SNLI-VE	CRIC	AOKVQA
ViLT	Question	**76.46**	**72.99**	**24.01**
	1 Fact + Question	75.60	71.38	13.45
	2 Facts + Question	69.88	71.64	13.45
	3 Facts + Question	65.44	62.00	11.38
	5 Facts + Question	64.62	38.30	13.25
VisualBERT	Question	**74.48**	**62.60**	**23.6**
	1 Fact + Question	74.34	61.00	22.0
	2 Facts + Question	73.99	20.25	20.0
	3 Facts + Question	73.98	19.00	19.0
	5 Facts + Question	73.62	19.00	19.0
FLAVA	Question	**79.93**	**73.11**	**33.07**
	1 Fact + Question	79.59	71.36	32.68
	2 Facts + Question	79.76	31.51	31.51
	3 Facts + Question	79.81	31.31	34.48
	5 Facts + Question	78.72	25.84	25.84

Table 10: Performance on varying the number of facts retrieved by finetuned ColBERTv2 concatenated with the query.

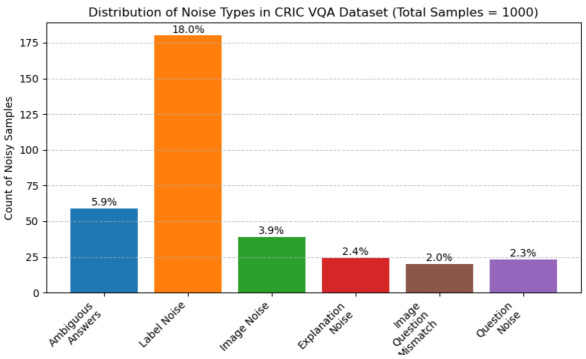

Figure 4: Noise Distribution of CRIC. We checked 1000 random samples, out of which 175 samples had label noise and others had the above distribution.

tion. For example, for Fig. 7b, 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 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.

G Noise

Noise in Datasets Figure 6 presents noisy samples from our chosen dataset that contribute to performance degradation. Through manual analysis of a substantial portion of the dataset, we identified issues such as noisy labels, ambiguous answers, and unclear questions. Below, we highlight some noteworthy examples illustrating these noise types.

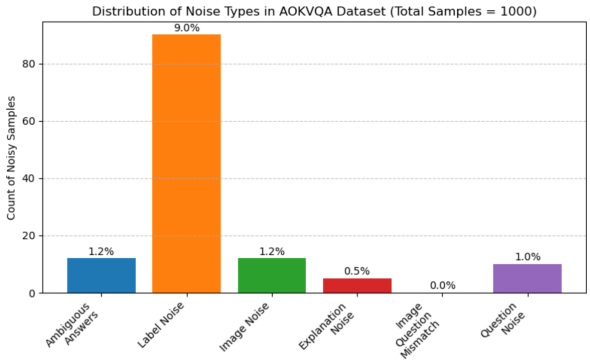

Figure 5: Noise Distribution of AOKVQA. We checked 1000 random samples, out of which around 90 samples had label noise, and others had the above distribution, which is significantly better than CRIC.

Additionally, Figs. 4 and 5 categorise the various types of noise present in the CRIC dataset.

Noise Ablations of e-SNLI-VE and AOKVQA Here we present the effect of using Symmetric Cross Entropy Loss and Generalised Cross Entropy Loss when applied to e-SNLI-VE and AOKVQA datasets in Tables 11&12 respectively. Other than FLAVA, both SCE and GCE+CE combinations provide noticeable gains. For e-SNLI-VE, however, such gains are not prominent. Also, we present the noise distribution of CRIC and AOKVQA, analysed by 3 independent raters who are computer science graduates, two of whom are authors of this paper.

Architecture	Type	CE	SCE	GCE + CE
ViLT	Type-5	78.46	77.75	**78.57**
VisualBERT	Type-5	78.83	77.23	**78.95**
FLAVA	Type-5	81.54	80.47	**82.05**

Table 11: Comparison of performance (in percentages) for different architectures using various loss functions for e-SNLI-VE.

Architecture	Type	CE	SCE	GCE + CE
ViLT	Type-5	28.15	**33.45**	32.33
VisualBERT	Type-5	35.40	**40.12**	38.51
FLAVA	Type-5	**47.85**	47.73	46.45

Table 12: Comparison of performance (in percentages) for different architectures using various loss functions for AOKVQA.

H Explanation Generation

Examples for Type-0, Type-1, Type-5 Explanations Here we provide more examples of the different **Types** of explanation produced by Llama-3.1-8B index (a - g). Of the 8 examples presented, only in (c) is the **Type-5** explanation incorrect. Even in that case, the explanation is still rationally correct because there is indeed no “object" on the aeroplane– 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.10. Although consistent, the **Type-1** explanations are nearly never precise enough to enable the downstream VLM to identify the correct answer.

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 tasks, it failed to generate a well-reasoned explanation concerning the context of its image and the questions supplied to it. To ensure the quality of data to finetune Florence, we used a training split of the VQAe dataset comprising 181K data instances. We used VQAe (Li et al., 2018) because it closely resembles the type of explanations that CRIC and e-SNLI-VE have. We report the Florence finetuned performance on CRIC in 1. In 9 we show the difference between explanations generated by the Florence vs LLAMA-3.1-8B **(Type 5)**.

LLM Prompts for Explanation Generation: LLaMA-3.1-8B & GPT-3.5 We feed GPT-3.5 and LLaMA-3.1-8B with Image captions, Objects Detected, and Retrieved Facts from OMCS, then ask it to analyse and generate a 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 not to produce the forbidden words. For instance, terms like "image description" and "caption" 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 explicitly

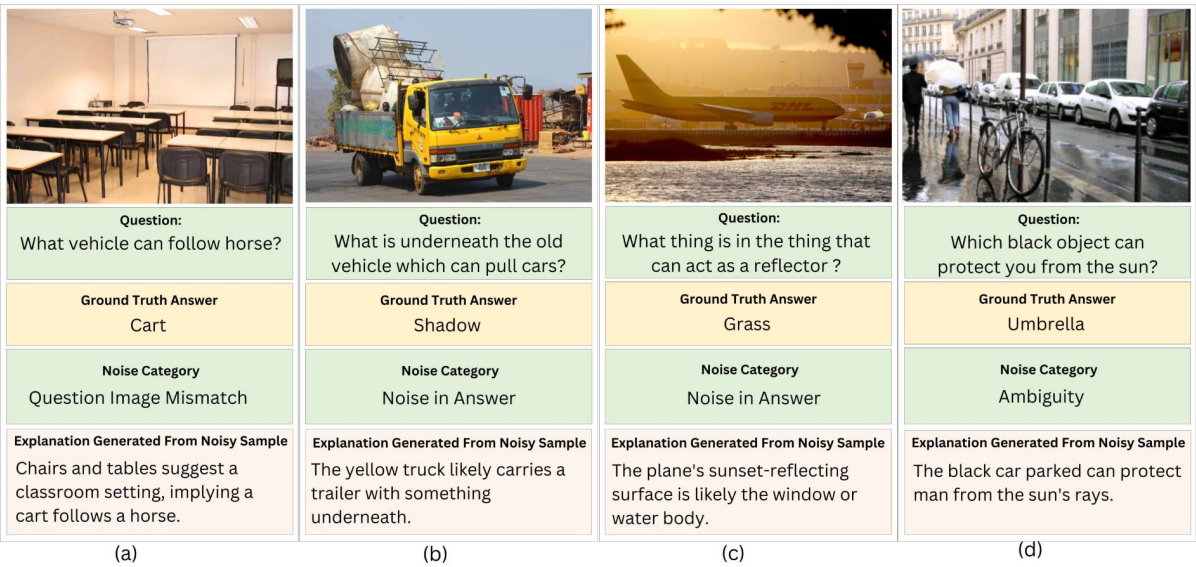

Figure 6: Examples illustrating the **noisy samples** from the datasets. where *Noise Category* is the type of noise we encountered during manual analysis, *Explanation Generated From Noisy Sample* is the LLaMA-3.1-8B explanation generated as a result of the noise for each sample, which degrades the overall performance.

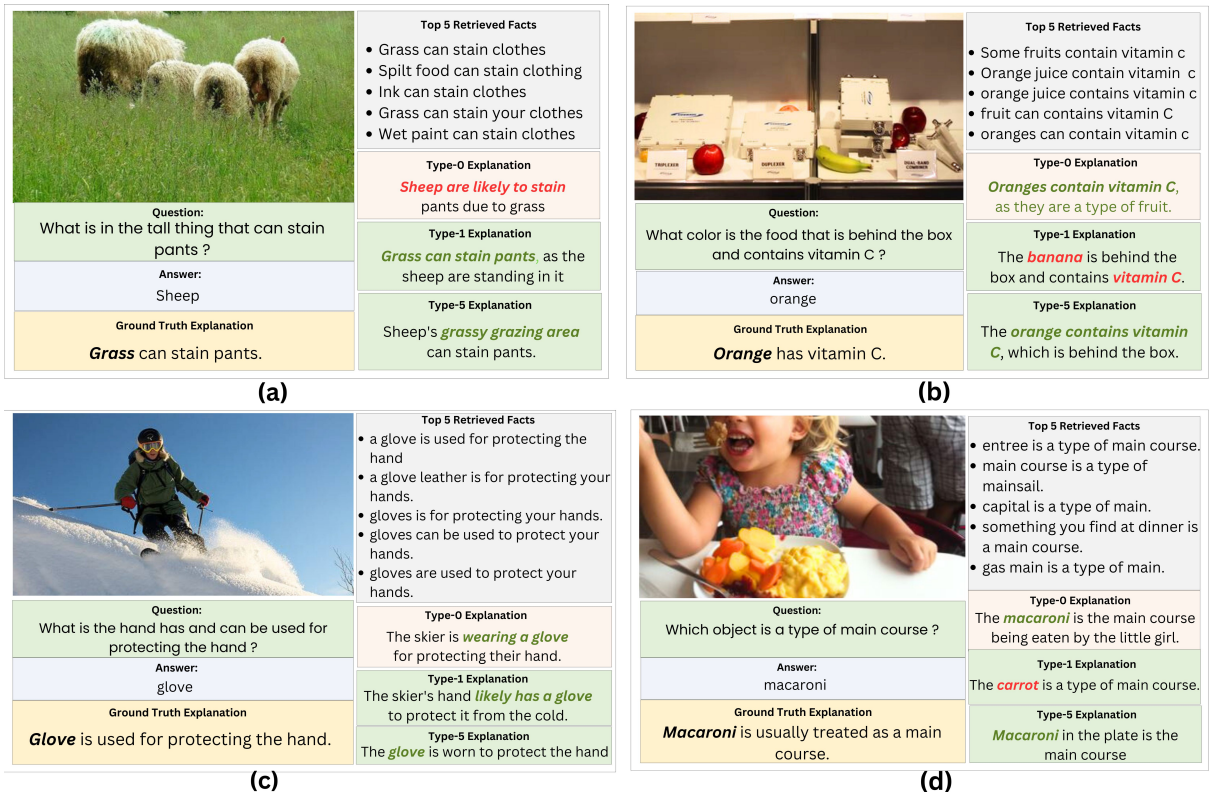

Figure 7: Examples from the CRIC dataset showing the difference between the Ground Truth, Type-0, Type-1 and Type-5 explanations generated by the LLama-3.1-8B. In all four cases, the **Type-5** explanation is the most rational and contextually relevant. **Type-0** and **Type-1** explanations are often hallucinated (to align the explanation with the label) or do not make rational sense. More examples are in the Appendix.

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 lower the performance the model delivers.

What We Define as Hallucinations and Their Constituents In our work, we used the term *"Hallucination"* repeatedly to indicate any content

Figure 8: More examples of **Type-0**, **Type-1**, and **Type-5** generated explanations.

in an LLM-generated explanation that is not supported by and in some cases directly contradicts the information in the question, the image, or any retrieved knowledge provided as context. In other

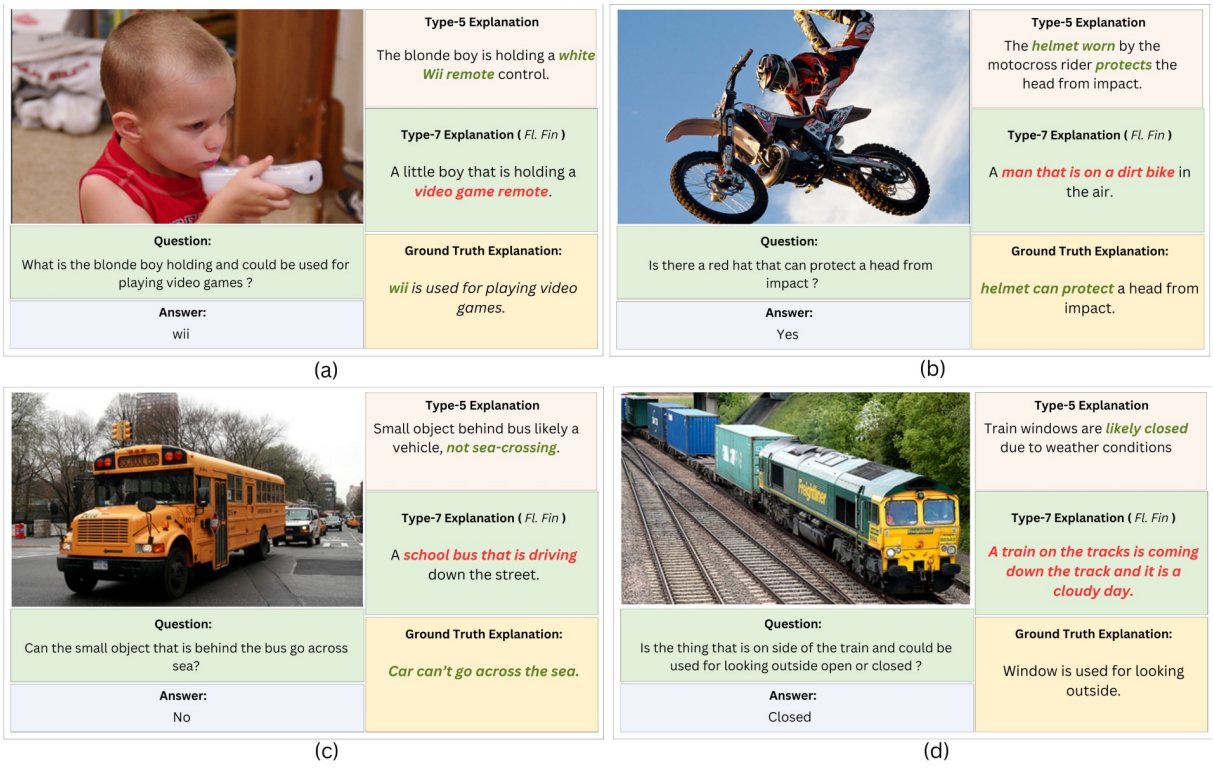

Figure 9: Examples showing difference between the LLAMA-3.1-8B generated **Type 5** explanation and Florence Finetuned Generated explanations **Type 7**

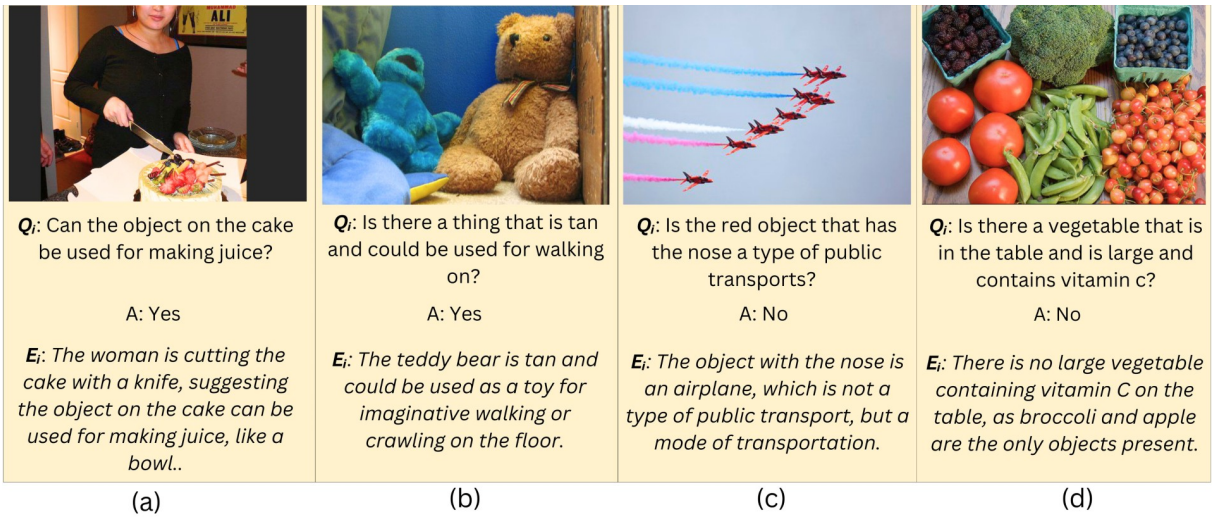

Figure 10: We report some instances where the LLM generates inaccurate **Type-0** explanations (where ground truth label was used). The visual and textual disconnects highlight the challenges LLMs face in reasoning, leading to 'hallucinations' that stray away from the visual context in the image.

words, whenever the generated text introduces details (objects, events, attributes, etc.) that do not appear in the image or are not logically inferable from the question and any external knowledge, we categorise those extraneous or fabricated details as hallucinations.

A typical example would be in Fig. 10d when the explanation states that the image contains *"apples"* even though the visual data and object detections show *"no apples"*. This also includes instances where the explanation confidently describes relationships or properties that are not depicted (e.g., *"the dog is wearing a red collar"* when no collar is visible). These are **hallucinated** elements because they are neither prompted by the data nor verifiable through the available context.

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**
AOKVQA	Explanations	Train	**49.29**	**27.86**	**39.88**	**29.55**	**23.45**	**65.31**
	Retrieved Fact	Train	0.3263	8.05	19.47	9.59	**5.96**	43.86
	Explanations	Val	**50.78**	**30.24**	**41.25**	**31.24**	**25.43**	**66.52**
	Retrieved Fact	Val	33.56	8.81	20.32	10.15	6.27	45.09

Table 13: 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.

Generating Type-5 Explanations with Various Llama Versions Since **Type-5** explanations are the ones that worked best for our use case, as seen in Table 1 and Table 2, we tried generating them with smaller models from the Llama family, aside from our primary Llama-3.1-8B model. We used Llama-3.2-1B & 3B for this. We report the scores in Table 14.

Ver	B-1	B-2	R-1	R-2	R-L	C
8B	**31.08**	**22.77**	**46.30**	**24.63**	**42.78**	**63.08**
3B	27.34	19.96	44.54	23.31	41.40	61.03
1B	23.10	16.76	39.60	20.07	35.60	55.87

Table 14: Quality of Type-5 explanations generated with different Llama versions.

AOKVQA Scores for Various Type-K Explanations We initially reported that deriving the Type-5 explanation was most beneficial to our CRIC dataset task. Table 15 reports the same score for AOKVQA, indicating that Type-5 explanations perform the best irrespective of the dataset, both in terms of capturing the visual context or guiding the VQA models for better predictions.

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 tasks from the knowledge bases. We have used a dual NVIDIA-A40 GPU system, each with 46 GB of memory, for all the experiments.

Given an image description, the task is to generate an explanation based on the following information.

Traditional-Caption-(TC): `traditionalcaption`$_I$.
Dense Caption (DC): `densecaption`$_I$.
Region Caption (RC): `regioncaption`$_I$.
Objects (O): `objects`$_I$.
Question (Q): Q.
Retrieved Facts (RF): $F_{I,Q}$.

You strictly need to produce a 15-20 word single-line explanation to help VQA models derive conclusions and nothing else. Forbidden words: **"image description", "captions"**.

Type	Explanation	B-1	B-2	B-3	R-1	R-2	R-L	Cosine
Type-2	DC + Q + $F_{I,Q}$	40.38	29.92	23.92	49.38	28.16	44.88	65.98
Type-3	RC + Q + $F_{I,Q}$	34.76	25.94	20.66	43.27	24.50	40.07	50.88
Type-4	RC + O + Q + $F_{I,Q}$	38.09	27.86	21.99	46.55	25.50	42.56	59.93
Type-5	DC + RC + O + Q + $F_{I,Q}$	**41.15**	**31.34**	**25.43**	**50.78**	**30.24**	**46.54**	**66.52**
Type-6	DC + RC + O + Q	39.13	28.84	22.96	48.52	27.23	43.90	64.95
Type-7	I + $< TP >$	17.49	7.51	4.48	20.45	2.57	16.78	28.08

Table 15: Results from the val split of AOKVQA showing a comparison of BLEU, ROUGE, and Cosine Scores (w.r.t. the GT explanation) for various types of explanations.

Architecture	LR	Batch Size	#Epoch	#Hours
ViLT	5e-5	32	6	15
VisualBERT	4e-5	32	6	14
FLAVA	4e-5	32	6	14.5
ColBERTv2	1e-4	32	4	3

Table 16: Hyperparameters for Finetuning tasks