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

Aritra Dutta[1], Swapnanil Mukherjee[2], Deepanway Ghosal[3], Somak Aditya[1]

[1]IIT Kharagpur [2]Ashoka University [3]Independent Researcher

The Problem: sVLMs Lack Commonsense

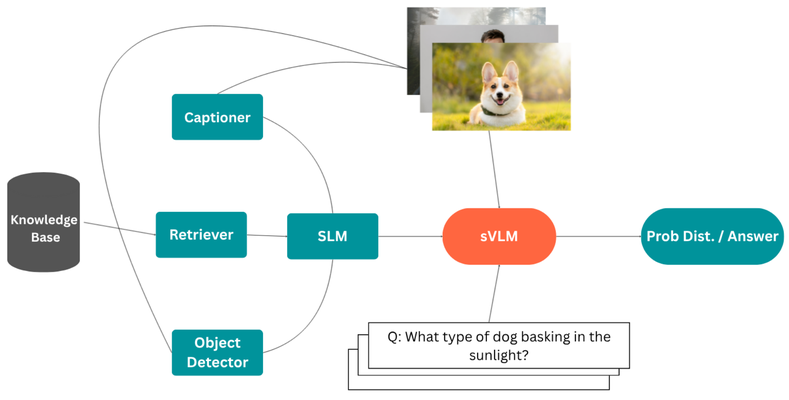
Small Vision-Language Models (sVLMs) often fail on questions that require external, real-world knowledge not present in the image.

Models like ViLT, VisualBERT, and FLAVA struggle to answer questions requiring reasoning about object properties, functions, or real-world scenarios not explicitly shown.

Figure: **Q: Is there a place that is blue and is liquid ?** An example VQA task requiring external commonsense knowledge: identifying that water is a "Liquid" state of matter and mostly "Blue" in colour in the natural habitats.

Figure: **Q: What place of the mountain is dangerous?**. The commonsense knowledge needed is: identifying that the shown "cliff" is a small rock surface, and it can pose a potential threat to life.

The Solution: NLKI Plug-and-Play Framework



Our framework enhances sVLMs by integrating knowledge in a modular, three-step process:

- **Retriever:** A fine-tuned **ColBERTv2** retrieves the most relevant commonsense facts from a knowledge base.
- **Context Generation:** A captioner (**Florence-2-large**) and object detector (**YOLOv8**) provide rich visual details.
- **LLM Explainer: Llama-3.1-8B** synthesises the visual context and retrieved facts into a concise, natural language explanation.

Results: NLKI Outperforms Baselines and Larger Models

Our NLKI framework significantly boosts the performance of all sVLMs. The combination of Type-5 explanations and noise-robust training yields best results, allowing models like FLAVA to outperform much larger generative models.

Architecture	Knowledge	e-SNLI-VE	CRIC	AOKVQA
ViLT	✗	76.46	72.99	24.01
Concat (ViLT)	Type 5	78.46	74.95	28.15
Concat (ViLT)*+CE	Type 5	**78.57**	**76.98**	**33.45**
VisualBERT	✗	74.48	62.60	31.59
Concat (VB)	Type 5	78.83	64.69	35.40
Concat (VB)*+CE	Type 5	**78.95**	**67.15**	**40.12**
FLAVA	✗	79.93	73.11	33.71
Concat (FLAVA)	Type 5	81.54	75.02	37.45
Concat (FLAVA)*+CE	Type 5	**82.05**	**77.85**	**47.85**

Table: Accuracy comparison across architectures and knowledge types. *+CE denotes noise-robust training.

Finding 1: Small Models Beat Giants

NLKI enables our 240M-parameter FLAVA model to match or outperform models 10x its size.

- On AOKVQA, NLKI boosts FLAVA's accuracy from **33.07% to 47.85%** (a +14.8% improvement).
- This performance surpasses larger models like Qwen2-VL (**41.90%**) and SmolVLM (**33.89%**) with significantly less compute.

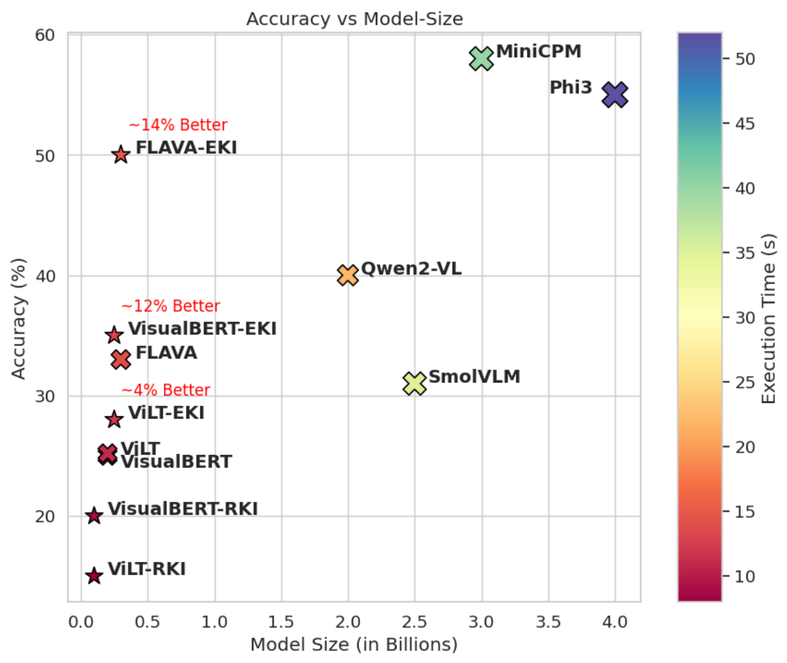


Figure: NLKI-powered sVLMs (top-left) achieve high accuracy at a fraction of the size and execution time of larger models.

Finding 2: Better Prompts, Better Explanations

Our "Type-5" prompt, which grounds the LLM in a rich visual context, is the most effective at generating accurate explanations and reducing hallucinations.

Compared to other prompt strategies, our Type-5 variant achieves the highest similarity to ground-truth explanations (e.g., **60.17 Cosine Similarity** on CRIC).

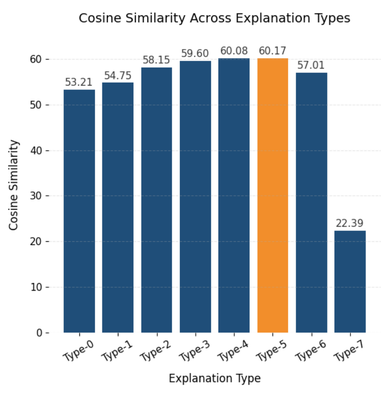


Figure: (Left) Quantitative results showing Type-5's superior similarity scores. (Right) Qualitative example. **Q:** *Is there a vegetable that is large and contains vitamin C?* **Type-5 Expl:** *The large head of broccoli in the centre is a vegetable known to be rich in vitamin C.*

Finding 3: Tackling Noisy Data

Commonsense VQA datasets contain significant label noise (∼18% in CRIC). Using a noise-robust loss function like Symmetric Cross-Entropy (SCE) is crucial for stable training.

On the CRIC dataset, SCE provides a vital **+2.8%** accuracy gain for FLAVA compared to the standard training baseline.

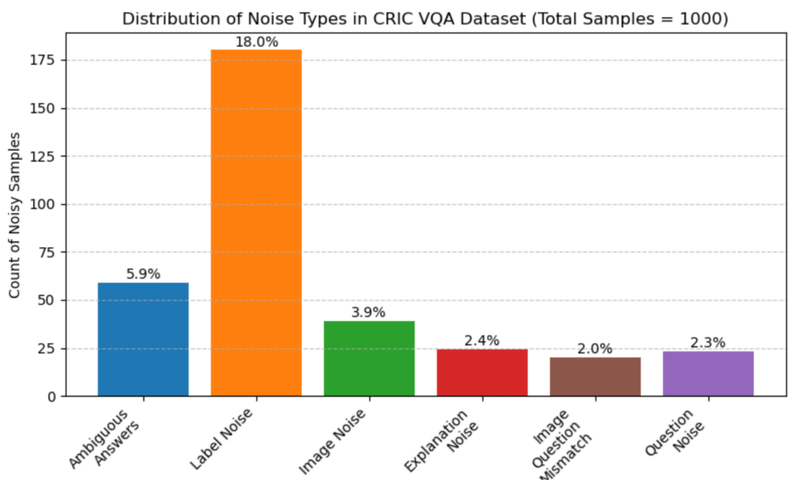


Figure: Distribution of noise types found in a 1000-sample audit of the CRIC dataset, with label noise being the most prevalent issue.

Inference Time Latency

The total pipeline latency is 1.32s per sample when run sequentially. Most of the cost comes from captioning and explanation generation. The pipeline can be optimised by running components concurrently or offline, making NLKI lightweight and deployable.

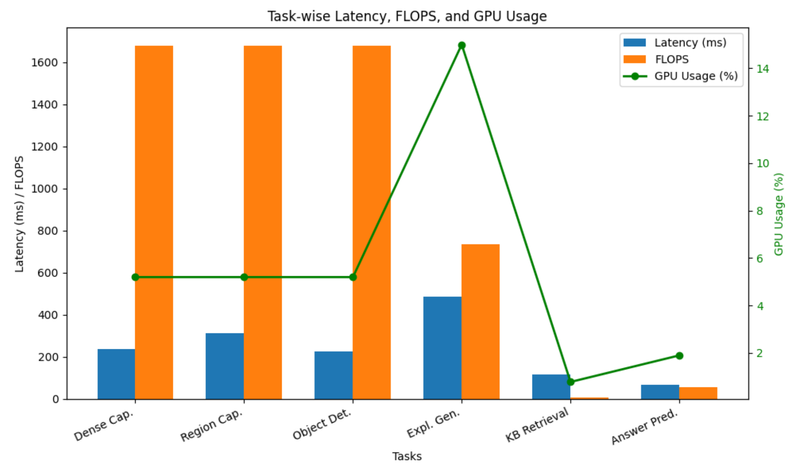


Figure: Breakdown of latency, FLOPs, and GPU usage across NLKI pipeline stages.

Conclusions

- **NLKI** is a lightweight, effective framework for commonsense VQA.
- **Rich visual context** in prompts is key for high-quality LLM explanations.
- **Noise-robust losses** are essential for stable training on real-world datasets.
- **The bottom line:** With NLKI, 250M-parameter models can rival multi-billion parameter giants.

Acknowledgement

We gratefully acknowledge the support of the DGX Cluster systems at IIT Kharagpur and the Science and Engineering Research Board (SERB), India, for making this work possible.