FilterRAG: Zero-Shot Informed Retrieval-Augmented Generation to Mitigate Hallucinations in VQA

Nobin Sarwar
University of Maryland, Baltimore County

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Background and motivations

Goal:

Address hallucinations and OOD vulnerability in knowledge intensive VQA

Question: What is the name of the items the hot dog are topped with?
Ground Truth: condiment, onion relish, vegetable, relish

Question: What sport can you use this for?
Ground Truth: race, motocross, ride

Fig. Two **OK-VQA** samples that depend on external knowledge about food and sports.

Motivations and challenges:

- Calibrate combined confidence from retrieval and generation for safe filtering.
- Map hallucination triggers under knowledge heavy questions and weak retrieval.
- Test OOD robustness across controlled category splits.

Contributions:

- **FilterRAG:** A retrieval-augmented approach that grounds VQA responses in external knowledge.
- **Zero-shot Learning:** Enhancing retrieval and reducing hallucinations in OOD scenarios.
- **Comprehensive Evaluation:** Evaluation on the OK-VQA dataset, demonstrating robustness and reliability for knowledge-intensive tasks.

FilterRAG framework

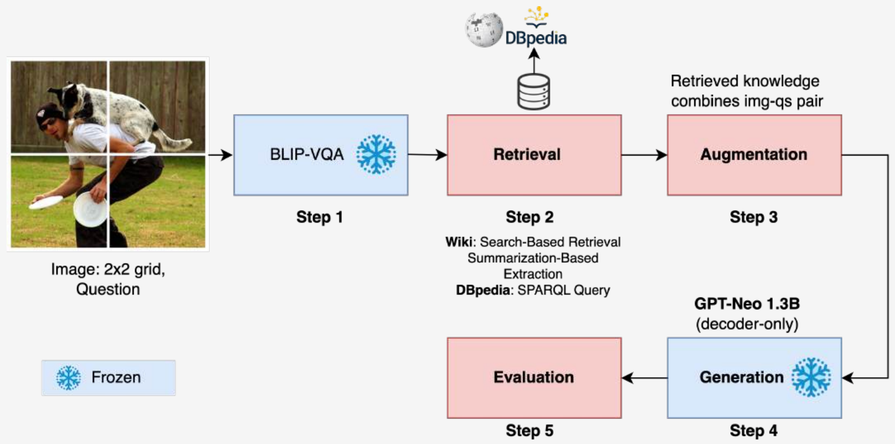

Fig. Overview of **FilterRAG** architecture: A step-by-step process integrating frozen **BLIP-VQA** with **Retrieval Augmented Generation (RAG)**. The system retrieves knowledge from Wikipedia and DBpedia, augments image-question pairs, and uses frozen **GPT-Neo 1.3B** to generate answers.

Quantitative evaluation on FilterRAG

Method	External Knowledge Sources	Acc. (%)
BAN + AN	Wikipedia	25.61
ConceptBERT	ConceptNet	33.66
KRISP	Wikipedia + ConceptNet	38.35
Vis-DPR	Google Search	39.2
MAVEx	Wikipedia + ConceptNet + Google Images	41.37
PICa-Full	Frozen GPT-3 (175B)	48.0
KAT (Ensemble)	Wikipedia + Frozen GPT-3 (175B)	54.41
FilterRAG (Ours)	Wikipedia + DBpedia (**Frozen** BLIP-VQA + GPT-Neo 1.3B)	**36.5**

Tab. Performance comparison of SOTA methods on the **OK-VQA** dataset

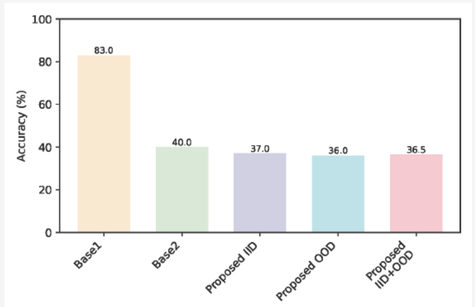

Fig. Comparison of model **accuracy** across settings

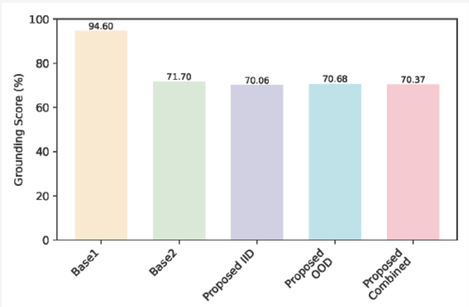

Fig. **Grounding score c**omparison across methods

Ablation study

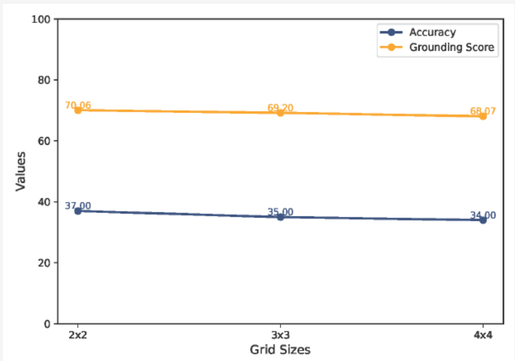

Fig. Effect of grid sizes on **accuracy** and **grounding score**

Qualitative evaluation

Question: A center affixed unit like this one in a kitchen is called a what?
Predicted Answer: island
Ground Truth: island
Setting: In Domain

Question: What do they call running around the bases on a single hit?
Predicted Answer: home run
Ground Truth: homerun, home run
Setting: Out-of-Distribution

Question: What is the name of the board he is on?
Predicted Answer: surfboard
Ground Truth: surf board, surfboard, surf
Setting: Out-of-Distribution

Question: What does this grow from?
Predicted Answer: flowers
Ground Truth: ground, plant, hibiscus plant stem, root
Setting: In Domain
Error: Wrong prediction

Question: What type of bike is on the ground?
Predicted Answer: dirt bike
Ground Truth: bmx, bicycle, 10 speed
Setting: Out-of-Distribution
Error: Wrong prediction

Question: Why is this plugged in?
Predicted Answer: plug
Ground Truth: charge, to have power and work, power, outlet
Setting: Out-of-Distribution
Error: Wrong prediction

Takeaways

- Retrieval grounding reduces hallucinations
- External retrieval improves OOD robustness
- Balanced image grids improve grounding and accuracy

Paper

Author website