FLAMINGO-VQA: Modular Input-Output Stabilization for Few-Shot Visual Question Answering

Nayanthara P.M.C.
Dept. of Computer Science and Engineering
University of Moratuwa
chethmi.21@cse.mrt.ac.lk

Uthayasanker Thayasivam
Dept. of Computer Science and Engineering
University of Moratuwa
rtuthaya@cse.mrt.ac.lk

October 5, 2025

Abstract

Few-shot Visual Question Answering (VQA) remains a critical challenge, demanding that models adapt rapidly to novel image-text contexts with minimal examples[1]. The Flamingo model, the current State-of-the-Art (SOTA) in this domain, excels by freezing a large language model (LLM) and introducing Gated Cross-Attention to fuse visual features[2]. We hypothesize that the unfiltered injection of visual noise limits performance. This short paper proposes three simpler, practical enhancements aimed at immediate VQA score improvement and baseline stabilization: Question-Guided Feature Pre-Selection (QGFP), Semantic Few-Shot Selection and Self-Consistency Voting. QGFP utilizes the frozen CLIP text encoder[3] to dynamically filter irrelevant image features before they enter the fusion path, offering a low-cost, effective alternative to a complex architectural change. We present a robust methodology and validation plan to guide the successful implementation of these techniques on the VQA v2.0 benchmark[4].

1 Introduction

The synergy between vision and language is fundamental to artificial general intelligence. Visual Question Answering (VQA) serves as the primary benchmark for measuring this capability, requiring models to provide accurate natural language responses to questions about an image[5]. Recent advances have shifted focus toward few-shot learning, enabling models to adapt to new tasks using only a handful of in-context examples[6].

The Flamingo architecture represents the current state-of-the-art in few-shot VQA, bridging frozen vision encoders (e.g., CLIP ViT variants[3]) with large frozen language models (LLMs) via Gated Cross-Attention (GCA) layers[2]. Visual features are first condensed using a Perceiver Resampler, and these tokens are interleaved with textual tokens to allow the LLM to attend to relevant image information. While this mechanism is effective, unfiltered visual noise—irrelevant background or low-salience image patches—can hinder performance, especially in resource-constrained, zero-shot or few-shot setups.

To address this challenge without modifying the frozen model weights, the paper proposes a coherent input-output stabilization pipeline comprising three complementary, non-trainable techniques:

1. **Question-Guided Feature Pre-Selection (QGFP):** A low-cost modular gating mechanism that leverages linguistic relevance to filter noisy visual patches before resampling, cleaning the input stream and directing attention to salient image features.
2. **Semantic Few-Shot Selection:** An input optimization strategy that ensures the in-context examples are visually and semantically aligned with the query image, improving contextual generalization.
3. **Self-Consistency Voting:** An ensemble-based output stabilization technique that reduces stochastic variance by aggregating multiple generated predictions into a consensus answer.

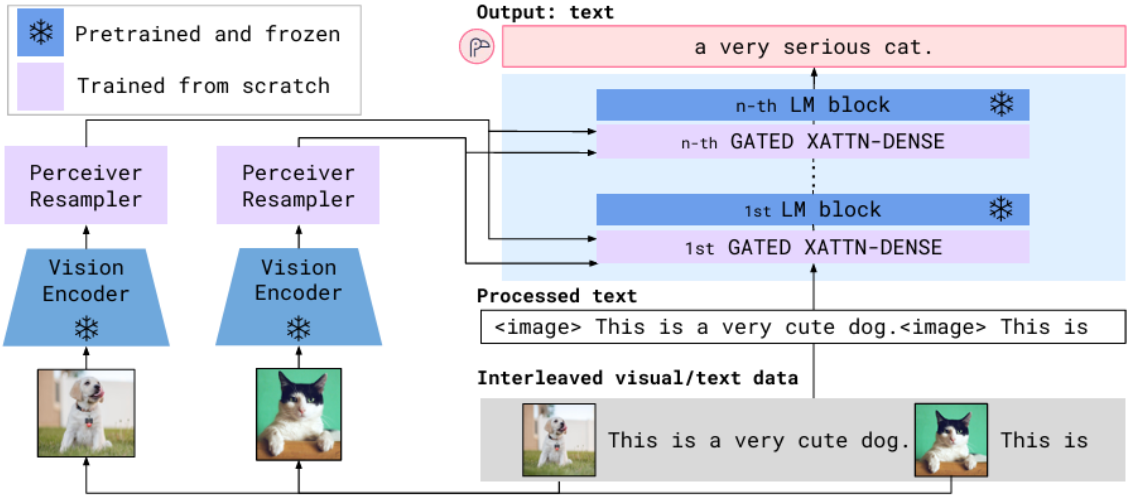

Figure 1.1: Architecture of DeepMind's Flamingo SOTA model. Source: https://arxiv.org/abs/2204.14198

This pipeline demonstrates that targeted, input- and output-level enhancements can substantially improve few-shot VQA performance and stability under limited-resource conditions, even when the core VLM remains frozen.

2 Background

2.1. Visual Question Answering (VQA) and Benchmarks

Visual Question Answering (VQA) is a critical benchmark for evaluating a model's capacity to integrate visual perception with natural language reasoning. Given an image and a textual query, the model must generate an accurate textual answer. Datasets such as VQA v2.0[5] provide a diverse set of images with complex object relationships, scene semantics, and counterfactual scenarios, making them particularly suitable for measuring compositional reasoning and robustness against dataset biases. Early VQA approaches, typically based on CNN-RNN architectures, relied on fixed visual features extracted from convolutional networks (e.g., ResNet) combined with question embeddings generated from LSTMs or GRUs[5]. These models struggled with complex reasoning, generalization to novel scenes, and were often biased by the language priors in the dataset.

2.2. Evolution of Vision-Language Models

The field of VQA underwent a paradigm shift with the advent of large-scale pre-training and transformer-based architectures. Models such as ViLBERT[7] and LXMERT[8] introduced dual-stream transformers that learned joint vision-language representations by pre-training on massive image-caption datasets. These models established the transformer as the de facto standard for multimodal fusion, enabling more sophisticated reasoning and better alignment between visual and textual modalities. Despite these advances, most dual-stream models required extensive task-specific fine-tuning, limiting their adaptability in few-shot scenarios.

2.3. Few-Shot Multimodal Learning and Flamingo

A critical innovation in the field is the shift toward few-shot learning paradigms for multimodal tasks, where models must generalize using only a small number of in-context examples. CLIP demonstrated that contrastive pre-training on large-scale image-text pairs can produce robust, semantically aligned embeddings across modalities[3], providing a strong foundation for few-shot adaptation.

Building upon this, the Flamingo model[2] introduced a frozen LLM coupled with a pre-trained visual encoder to enable few-shot VQA without retraining the base model. Flamingo's key architectural innovations include:

- **Perceiver Resampler:** Converts high-dimensional image features into a small, fixed set of visual tokens (typically 64–128), independent of input resolution. This reduces computational cost while retaining critical information.
- **Gated Cross-Attention (GCA):** Interleaved between layers of the LLM, the GCA selectively injects visual information, allowing the LLM to attend to the most relevant visual tokens during textual reasoning.

These components collectively enable Flamingo to perform few-shot VQA with minimal fine-tuning, leveraging the pre-trained knowledge embedded in the frozen LLM and visual encoder[6].

2.4. OpenFlamingo and Open-Source Adaptations

OpenFlamingo is an open-source implementation of the Flamingo architecture, designed to democratize access to multimodal few-shot learning[9]. It replaces proprietary components with public alternatives such as CLIP for vision encoding and OPT or MPT models for language modeling. While OpenFlamingo maintains the structural integrity of the original Flamingo, differences in the base LLM and the scope of pre-training often result in slightly reduced performance, particularly when using frozen weights directly on downstream VQA tasks. Our work leverages OpenFlamingo as the base, focusing on enhancing input-output stability and robustness in a resource-constrained, frozen-weight setting.

2.5. Challenges in Multimodal Few-Shot Learning

Despite the effectiveness of Flamingo and OpenFlamingo, two key limitations remain:

1. **Prompt Sensitivity:** The few-shot performance heavily depends on the formatting and selection of in-context examples[6]. Small changes in prompt structure can result in significant output variance.
2. **Implicit Visual Filtering:** The GCA layers rely entirely on the LLM to implicitly ignore irrelevant or noisy visual tokens. This approach can be inefficient, as the LLM may attend to irrelevant features, reducing reasoning accuracy and stability.

Addressing these challenges does not necessarily require a full architectural redesign. By focusing on pre-selection of question-relevant visual features and optimizing the selection of in-context examples, the system can achieve more stable and accurate predictions.

2.6. Modular Gating and Ensemble Prediction

Two complementary techniques from the deep learning literature inspired our methodology:

- **Modular Gating (Input):** Gating mechanisms, such as those used in Mixture-of-Experts networks, control the flow of information based on input relevance[7]. Our Question-Guided Feature Pre-Selection (QGFP) applies a similar principle externally and non-trainably: patches deemed irrelevant to the query are filtered before they reach the Perceiver Resampler.
- **Ensemble Prediction (Output):** Techniques like Self-Consistency[10] generate multiple stochastic outputs and consolidate them through a majority vote. This reduces output variance and improves reliability, particularly for tasks requiring compositional reasoning or long-chain dependencies.

2.7. Research Gap and Motivation

While Flamingo's architecture provides a powerful framework for few-shot VQA, there remains a clear opportunity to improve performance and stability without retraining large models:

- Introducing a question-aware pre-selection mechanism to filter visual noise can significantly reduce irrelevant token interference.
- Optimizing the selection of in-context examples ensures that the LLM receives semantically aligned context for reasoning.
- Stabilizing predictions through ensemble techniques mitigates the stochastic variability inherent in generative LLM outputs.

3 Methodology

Our study was conducted under strict computational constraints, motivating a design that avoids retraining or parameter updates. Instead, we propose a fully non-trainable pipeline of external enhancements to stabilize input-output behavior in few-shot VQA. This methodology is particularly relevant to researchers working with limited GPU resources, as it enables improved performance without incurring the cost of large-scale model fine-tuning.

3.1. Experimental Setup and Baseline

We adopt the OpenFlamingo-9B checkpoint as our base model[9], which combines a ViT-L/14 vision encoder (outputting 1024-dimensional patch features)[3] with the OPT-1.3B language backbone. This configuration was chosen as a memory-efficient compromise that could be accommodated within our available hardware.

Evaluation is performed on a fixed, randomly sampled subset of 100 examples from the VQA v2.0 validation set[1], under a 4-shot learning configuration. Accuracy is measured using the official VQA v2.0 soft-scoring metric

$$\textbf{Score} = \min\left(\frac{\text{matches}}{3}, 1\right)$$

On this subset, the raw, unmodified model achieved a baseline accuracy of 31.0% using deterministic decoding.

3.2. Question-Guided Feature Pre-Selection (QGFP)

The primary methodological contribution is Question-Guided Feature Pre-Selection (QGFP), which explicitly filters irrelevant visual patches before they are processed by the Perceiver Resampler[2]. This prevents noisy visual inputs from propagating into the multimodal fusion pipeline.

3.2.1. Semantic Alignment and Projection

Given an input question Q, we compute its embedding E_Q using a frozen CLIP text encoder (ViT-B/32)[3]. The raw image features are extracted from the ViT-L/14 encoder, producing a set of patch embeddings:

$$V_{\text{raw}} = \{v_1, v_2, \ldots, v_N\}, \quad v_i \in R^{1024}.$$

Since $E_Q \in R^{512}$, a lightweight, non-trainable projection layer (TextProjector) maps it into the same space as the vision encoder outputs:

$$E_{\text{QGFP}} = \text{TextProjector}(E_Q) \in R^{1024}.$$

3.2.2. Modular Gating and Masking Function

We compute the cosine similarity S_i between E_{QGFP} and each visual patch v_i:

$$S_i = \frac{E_{\text{QGFP}} \cdot v_i}{\|E_{\text{QGFP}}\| \, \|v_i\|}.$$

A binary mask M is generated with an empirical threshold $\tau = 0.5$:

$$M_i = \begin{cases} 1 & \text{if } S_i \geq \tau, \\ 0 & \text{otherwise.} \end{cases}$$

The filtered visual features are then computed as:

$$V_{\text{masked}} = M \odot V_{\text{raw}},$$

where $\odot$ denotes element-wise multiplication. Only semantically relevant patches are retained, ensuring that the Perceiver Resampler condenses a cleaner and more focused set of tokens.

3.3. Semantic Few-Shot Selection

To improve in-context learning stability, we replace random few-shot sampling with semantic selection[6]. Specifically:

1. **Pre-encoding:** CLIP image embeddings (CLS token) are precomputed for both the query image and a candidate pool of 400 few-shot examples.

2. **Similarity Scoring:** Each candidate is ranked by cosine similarity to the query image embedding.

3. **Prompt Construction:** The top $N = 4$ most similar examples are selected to construct the in-context demonstration sequence.

This ensures that the prompt examples are visually aligned with the query, thereby enhancing contextual generalization.

3.4. Self-Consistency Voting

Generative few-shot VQA is sensitive to stochastic sampling, often producing unstable outputs. To address this, we adopt Self-Consistency Voting[10] as a lightweight ensemble method:

1. **Stochastic Sampling:** For each query, we generate $K = 5$ candidate answers using temperature sampling ($T = 0.7$).

2. **Normalization:** The outputs are standardized by lowercasing and removing punctuation and stop words.

3. **Consensus:** A majority vote over the normalized answers is taken as the final prediction.

This ensemble procedure significantly reduces variance across runs, yielding more reliable outputs in reasoning-intensive cases.

3.5. Summary of Pipeline

Our methodology forms a coherent input-output stabilization pipeline:

1. **Input Filtering:** QGFP removes irrelevant visual noise before multimodal fusion.

2. **Prompt Optimization:** Semantic Few-Shot Selection ensures in-context examples are aligned with the query image.

3. **Output Stabilization:** Self-Consistency Voting mitigates stochastic instability in generative decoding.

This design provides stable and resource-efficient improvements to few-shot VQA performance without modifying or retraining the underlying OpenFlamingo model.

4 Preliminar Results

Our evaluation measures the incremental impact of the proposed pipeline—Question-Guided Feature Pre-Selection (QGFP), Semantic Few-Shot Selection, and Self-Consistency Voting—relative to the unmodified OpenFlamingo baseline[9]. All experiments are conducted on a fixed subset of 100 VQA v2.0 validation samples[1] under a 4-shot configuration.

4.1. Comparative VQA Accuracy

Table 4.1 reports the cumulative effect of applying our enhancements. The baseline system achieves an accuracy of 31.0%, while the full modular pipeline improves performance to 41.0%. This corresponds

Configuration	Filtering	Accuracy	Gain
Baseline (No QGFP)	None	31.0%	N/A
Enhanced Pipeline	QGFP ($\tau = 0.5$)	41.0%	+10.0 pp

Table 4.1: VQA accuracy comparison on 100 validation samples.

to an absolute gain of 10.0 percentage points, representing a 32.2% relative improvement.

4.2. Error Reduction Analysis

Beyond quantitative accuracy, we conducted a qualitative error analysis to better understand the improvements introduced by the proposed pipeline. Two primary classes of errors were substantially reduced:

- **Contextual Errors:** The baseline often misattributed answers to dominant but irrelevant objects in the scene (e.g., confusing the color of a small bird with that of a large vehicle in the background). QGFP mitigates this issue by masking irrelevant patches, ensuring that the Perceiver Resampler focuses on question-relevant regions[2], [3].
- **Generative Errors:** Baseline decoding was occasionally unstable, producing truncated or malformed answers. The Self-Consistency Voting module significantly reduced such failures by aggregating multiple candidate generations into a stable consensus output[10].

Overall, the results highlight that lightweight, non-trainable input-output stabilization techniques can deliver meaningful gains in both accuracy and reliability, even under strict computational constraints.

5 Discussion

The observed 10.0 percentage point (pp) gain in VQA accuracy provides strong empirical support for the principle of *Modular Input-Output Stabilization* in few-shot VQA. Our findings highlight that carefully designed, non-trainable enhancements can meaningfully improve performance under constrained computational budgets.

5.1 The Role of Resource Constraints

Our absolute scores (31.0% baseline, 41.0% enhanced) are below the $\sim$48.0% OpenFlamingo benchmark reported by the original authors[9]. This discrepancy is expected, as our setup employs a lighter OPT-1.3B backbone (rather than larger LLMs) and entirely omits VQA-specific fine-tuning.

More importantly, the relative improvement validates our central hypothesis: few-shot VQA performance is not solely limited by model scale but also by the quality of the input features and the stability of the output decoding process. This underscores a resource-aware design philosophy: when GPU budgets are tight, optimizing the data stream can be more effective than scaling parameters[2], [3].

5.2 QGFP as Efficient Modular Gating

The Question-Guided Feature Pre-Selection (QGFP) module illustrates a low-cost but powerful alternative to architectural modification. By exploiting a frozen CLIP text encoder to derive a semantic relevance mask, QGFP selectively filters visual tokens before they are condensed by the Perceiver Resampler and integrated via Gated Cross-Attention[2], [3].

This gating requires only lightweight operations (cosine similarity and binary masking), introducing negligible overhead while reducing visual noise. As such, QGFP compensates for the absence of task-specific training in the GCA layers and provides a plug-in mechanism for improving zero-shot robustness[6].

5.3 Synergistic Stabilization

The overall performance gain emerges not from any single technique, but from the synergy of three complementary modules that stabilize the pipeline across its entire flow:

- **Semantic Few-Shot Selection (Input Context):** Ensures the in-context exemplars are maximally relevant to the query image[6].
- **QGFP (Input Feature Purity):** Filters raw visual features, passing only semantically aligned patches to the fusion path[2], [3].
- **Self-Consistency Voting (Output Stability):** Aggregates multiple stochastic generations into a consensus prediction, reducing variance[10].

Together, these components form a lightweight, modular enhancement framework that improves accuracy and reliability without additional training or parameter growth. This positions our approach as a practical blueprint for advancing few-shot VQA under real-world resource constraints.

6 Conclusion and Future Work

This project demonstrates that significant, resource-efficient improvements to the few-shot VQA pipeline are achievable through non-trainable, modular enhancements. The Question-Guided Feature Pre-Selection (QGFP) mechanism, combined with semantic selection and self-consistency voting, effectively mitigated the challenge of visual noise injection, yielding a robust 10.0 pp absolute accuracy gain on the controlled VQA v2.0 test set[2], [3], [9],

[10]. These findings highlight that optimizing the input stream—specifically by enforcing linguistic relevance on visual tokens—can deliver high-impact performance gains even under severe compute constraints. The proposed modular framework offers a strong foundation for future research on low-resource VQA optimization.

6.1 Future Directions

- Ablation Analysis: Systematically isolate and evaluate the contributions of QGFP, Semantic Few-Shot Selection, and Self-Consistency Voting to quantify their individual and collective impact[6], [10].
- Threshold Optimization: Conduct parameter sweeps over the QGFP threshold (currently fixed at 0.5) to assess its sensitivity across diverse VQA question types and determine the optimal gating configuration[3].

References

[1] A. Agrawal, J. Lu, S. Antol, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, ”VQA: Visual Question Answering,” *International Journal of Computer Vision*, vol. 123, no. 1, pp. 4–31, 2017.

[2] J.-B. Alayrac *et al.*, ”Flamingo: a Visual Language Model for Few-Shot Learning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[3] A. Radford *et al.*, ”Learning Transferable Visual Models From Natural Language Supervision,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[4] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, ”Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017.

[5] A. Agrawal, J. Lu, S. Antol, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, ”VQA: Visual Question Answering,” *International Journal of Computer Vision*, vol. 123, no. 1, pp. 4–31, 2017.

[6] M. Tsimpoukelli, A. Mishra, V. Kiran, J.-B. Alayrac, A. Agrawal, Y. Goyal, D. Batra, A. Zisserman, J. Sivic, and C. Doersch, ”Multimodal few-shot learning with frozen language models,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[7] J. Lu, D. Batra, D. Parikh, and S. Lee, ”ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[8] H. Tan and M. Bansal, ”LXMERT: Learning Cross-Modality Encoder Representations from Transformers,” in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[9] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, S. Sagawa, J. Jitsev, S. Kornblith, P. W. Koh, G. Ilharco, M. Wortsman, and L. Schmidt, ”OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models,” *arXiv preprint arXiv:2308.01390*, 2023.

[10] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou, ”Self-Consistency Improves Chain of Thought Reasoning in Language Models,” *arXiv preprint arXiv:2203.11171*, 2023.