Evaluating the Effectiveness of Question-Guided Captioning in Zero-Shot Visual Question Answering

By Alexis Jin (xj606)
And Aditya Maheshwari (am15247)
Demo: https://stream.nyu.edu/media/t/1_2g5tp71n

Overview of the Topic

1. Introduction to the model, why we chose it

Visual Question Answering (VQA) stands as a significant task within the realm of artificial intelligence, demanding a confluence of vision and language understanding to provide natural language answers to questions about images.[1] This task is particularly challenging in the zero-shot setting, where models are expected to perform without any specific training data for the VQA task itself. The inherent complexity arises from the necessity for a system to not only perceive and interpret visual content but also to comprehend and reason based on natural language queries.[1] Traditional approaches to VQA often necessitate the use of extensive human-annotated question-answer pairs for training, which can be a resource-intensive and time-consuming process.[4] Furthermore, many existing methods require substantial adaptation of pre-trained language models (PLMs) to effectively process visual information.[3] This adaptation typically involves additional training stages and the introduction of new network components, adding to the complexity and computational cost.[7]

In response to these limitations, the Plug-and-Play VQA (PNP-VQA) framework has emerged as a novel and modular approach to tackle the zero-shot VQA challenge.[1] This framework distinguishes itself by its ability to leverage the power of large pre-trained models without requiring any additional training specifically for the VQA task.[1] The core innovation of PNP-VQA lies in its strategic use of natural language and network interpretation as an intermediary representation to effectively connect pre-trained vision and language models.[1] By avoiding the need for extensive fine-tuning, PNP-VQA offers a potentially more efficient and adaptable solution for visual question answering, particularly in scenarios where labeled VQA data is scarce or non-existent. This report aims to provide a comprehensive analysis of the PNP-VQA framework, delving into its architecture, methodology, performance, advantages, limitations, and its position within the broader landscape of VQA research.

2. PNP-VQA: Architecture and Methodology

The PNP-VQA framework adopts a modular architecture, a design choice that provides significant flexibility and adaptability.[1] This modularity allows for the independent selection and combination of different pre-trained models, enabling the system to potentially evolve and improve as more advanced models become available.[3] The framework operates through a sequence of three primary modules: image question matching, image captioning, and question answering.[7]

The first crucial step in the PNP-VQA process involves generating question-guided informative image captions.[1] To achieve this, the framework first identifies the regions within the input image that are most relevant to the given question.[4] This identification is facilitated by a network interpretability technique, with GradCAM being a likely candidate.[4] GradCAM, a gradient-based localization method, allows the model to determine which parts of the image are most influential in processing the query, thereby enhancing the relevance of the subsequent captioning.[4] Once the question-relevant image regions are identified, the framework proceeds to generate captions based on these specific areas.[3] This caption generation is typically performed using a pre-trained vision-language model (PVLM) such as BLIP.[3] BLIP, having been trained on a vast dataset of image-caption pairs, possesses the capability to generate natural language descriptions of visual content.[4] The question guidance in this step is critical as it encourages the generation of captions that are not only descriptive of the image but also specifically tailored to the content of the question, thereby increasing the likelihood that these captions will contain valuable answer cues for the subsequent question answering stage.[7] This approach has been shown to potentially surpass the quality and relevance of generic or even human-written image captions for the purpose of VQA.[7]

The second key step in the PNP-VQA methodology is question answering, which is performed using a pre-trained language model (PLM).[1] The question-guided image captions generated in the previous step are passed to the PLM as contextual information, along with the original question.[1] By providing the PLM with visual context in a textual format, the framework leverages the language understanding and reasoning capabilities of the PLM to generate an answer to the question.[1] To effectively handle situations where multiple question-guided captions are generated, PNP-VQA often employs techniques like Fusion-in-Decoder (FiD).[1] FiD allows the model to process multiple captions in conjunction with the question, overcoming potential input length restrictions of the PLM and enabling the integration of information from various textual descriptions of the relevant visual content.[1] This two-step process, involving question-guided caption generation followed by question answering using a PLM, forms the core of the PNP-VQA framework's methodology for

tackling zero-shot visual question answering.

3. Advantages of PNP-VQA

The Plug-and-Play VQA (PNP-VQA) framework offers several notable advantages over existing Visual Question Answering (VQA) methods, particularly in the challenging zero-shot setting. One of its most significant benefits is that it requires **zero additional training of Pretrained Language Models (PLMs)**.[1] This is a substantial advantage because it circumvents the need for large-scale labeled VQA datasets, which can be expensive and time-consuming to acquire and annotate.[3] Instead, PNP-VQA leverages the general knowledge and language understanding capabilities that are already embedded within these PLMs from their pre-training on massive text corpora.[3] This training-free aspect makes PNP-VQA particularly appealing and practical for zero-shot learning scenarios where task-specific training data is often scarce or unavailable.[5]

Another key advantage of PNP-VQA is its **modular framework**.[1] This modularity provides a high degree of flexibility, allowing researchers and practitioners to choose and combine different state-of-the-art pre-trained models for each of the framework's components, such as image captioning and question answering.[7] This design enables the system to be easily adapted and potentially improved by simply swapping in more advanced pre-trained models as they become available, without the need for extensive retraining of the entire system.[7]

Furthermore, PNP-VQA effectively utilizes **natural language and network interpretation as an intermediate representation**.[1] This approach serves as a crucial bridge between the vision and language modalities, allowing pre-trained language models to process and reason about visual content without requiring direct multimodal training.[1] By converting visual information into natural language captions that are guided by the question, PNP-VQA creates a common ground that enables effective interaction between unimodal pre-trained models for the multimodal VQA task.[3]

Despite not undergoing any additional training for the VQA task, PNP-VQA has demonstrated **state-of-the-art performance in zero-shot VQA** on several challenging benchmark datasets, including VQAv2 and GQA.[1] This achievement highlights the effectiveness of the proposed plug-and-play approach in leveraging the knowledge and capabilities of existing pre-trained models for complex multimodal reasoning.[6]

Notably, PNP-VQA exhibits remarkable **efficiency**, as it has been shown to outperform significantly larger models, such as the 80 billion parameter Flamingo model, with a much smaller 11 billion parameter configuration on the VQAv2 dataset.[1] This demonstrates that the architectural design and methodology of PNP-VQA are highly effective in utilizing the model's parameters to achieve strong performance in zero-shot visual question answering.[1]

4. Performance Evaluation and Comparison

The performance of the Plug-and-Play VQA (PNP-VQA) framework has been evaluated on several standard Visual Question Answering (VQA) benchmark datasets, including VQAv2, GQA, and OK-VQA, demonstrating its capabilities in the zero-shot setting. On the **VQAv2 dataset**, PNP-VQA has shown promising results.[1] Notably, it surpasses end-to-end trained baselines on this dataset, indicating the effectiveness of its approach in leveraging pre-existing knowledge without task-specific training.[1] A significant achievement highlighted in the research is that the 11 billion parameter version of PNP-VQA outperforms the much larger 80 billion parameter Flamingo model by 8.5% on the VQAv2 dataset.[1] This result underscores the parameter efficiency of PNP-VQA. When compared to other zero-shot models on VQAv2, such as Frozen and FewVLM, PNP-VQA also demonstrates competitive performance.[1] Specifically, PNP-VQA achieved an accuracy of 63.3% on the VQA v2 validation set and 64.8% on the test-dev set.[15] While more recent models like BLIP-2 have achieved higher zero-shot accuracy on VQAv2 (e.g., 65.2%), PNP-VQA's performance remains highly competitive within the zero-shot VQA landscape [16]

On the **GQA dataset**, which is known for requiring more complex reasoning, PNP-VQA has also shown significant success.[1] In particular, with a 738 million parameter PLM, PNP-VQA achieves a 9.1% improvement on GQA compared to FewVLM, which utilizes a 740 million parameter PLM.[1] This comparison further highlights the efficiency of PNP-VQA in leveraging its parameters for the GQA task. The accuracy of PNP-VQA on the GQA test-dev set has been reported as 41.9%.[17]

Regarding the **OK-VQA dataset**, which focuses on questions requiring external knowledge, PNP-VQA has also been evaluated.[7] The reported accuracy for PNP-VQA on this dataset is 35.9%.[17] When compared with other zero-shot methods on OK-VQA, PNP-VQA's performance places it competitively among approaches that do not utilize additional training data.[19] For instance, comparisons with PICa and Img2LLM on OK-VQA indicate PNP-VQA's standing within the knowledge-based VQA domain.

To provide a consolidated view of PNP-VQA's performance, the following table

summarizes its accuracy on the key benchmark datasets in comparison to other notable models in the zero-shot setting:

Model	Dataset	Accuracy (%)	Parameter Count
PNP-VQA	VQAv2 val	63.3	11B
PNP-VQA	VQAv2 test-dev	64.8	11B
PNP-VQA	GQA test-dev	41.9	738M
PNP-VQA	OK-VQA	35.9	11B
Flamingo (zero-shot)	VQAv2 val	56.3	80B
BLIP-2 (zero-shot)	VQAv2 val	65.2	-
FewVLM (zero-shot)	VQAv2 val	47.7	740M
FewVLM (zero-shot)	GQA test-dev	32.8	740M

This table provides a clear overview of PNP-VQA's performance across different datasets and in comparison to other prominent zero-shot VQA models, highlighting its strong performance and efficiency.

5. Limitations and Challenges

Despite its notable achievements, the Plug-and-Play VQA (PNP-VQA) framework is not without limitations and challenges. One potential area of concern lies in its ability to handle **complex reasoning** tasks.[4] While PNP-VQA demonstrates strong overall zero-shot performance, tasks requiring intricate spatial reasoning, precise counting of objects, or multi-hop inference might still pose difficulties.[4] These types of reasoning often benefit from more direct processing of visual and textual information, rather than relying solely on the intermediate step of caption generation.

Another inherent limitation stems from the framework's **reliance on the quality and capabilities of the underlying pre-trained models**.[3] The accuracy of PNP-VQA is directly tied to how well the pre-trained vision-language model (PVLM) can generate

relevant and informative captions, and how effectively the pre-trained language model (PLM) can reason based on these captions and the question. If either of these pre-trained models has weaknesses or biases, it can negatively impact the overall performance of PNP-VQA.[7]

The **caption generation process** itself can also introduce limitations.[3] Converting the rich and nuanced information present in an image into a textual description can be a lossy process. Important visual details, spatial relationships between objects, or fine-grained attributes might not be fully captured in the generated captions, potentially hindering the PLM's ability to answer certain types of questions accurately.[8]

When comparing PNP-VQA's performance to **fully supervised VQA models**, a significant performance gap still exists.[3] While PNP-VQA excels in the zero-shot setting, models trained end-to-end on large datasets of image-question-answer pairs generally achieve higher accuracy across a broader range of VQA tasks.[4] This indicates a trade-off between the flexibility and training-free nature of PNP-VQA and the potentially higher accuracy achievable through supervised learning.

Furthermore, some research suggests that PNP-VQA might exhibit **lower performance compared to fine-tuned models with similar parameter counts**.[29] While PNP-VQA demonstrates impressive zero-shot capabilities, models that undergo task-specific fine-tuning on VQA datasets, even with comparable parameter sizes, might achieve better results due to the adaptation of their parameters to the specific nuances of the VQA task.[29]

6. Interpretability of Question-Guided Captioning

The question-guided captioning aspect of the Plug-and-Play VQA (PNP-VQA) framework offers a valuable contribution to the **interpretability** of the model's predictions.[1] By generating captions that are specifically relevant to the question being asked, PNP-VQA provides insights into the visual information that the model deems most important for arriving at an answer.[7] Examining these generated captions allows users to understand, to some extent, the model's focus and the rationale behind its response.[7]

The use of network interpretation techniques, such as GradCAM, further enhances the interpretability of PNP-VQA.[4] GradCAM highlights the specific regions within the image that most influence the generation of the question-guided captions and, consequently, the final answer.[4] This visual explanation of the model's attention allows

for a more transparent understanding of which parts of the image contributed most significantly to the model's decision-making process.[7]

Compared to **end-to-end trained VQA models**, which are often considered "black boxes" due to the difficulty in understanding their internal reasoning, PNP-VQA offers a higher degree of interpretability.[9] The modular design and the intermediate representations of question-guided captions and saliency maps (generated by GradCAM) provide a more transparent view of how the model processes the visual and textual inputs to produce an answer.[9]

PNP-VQA's approach aligns with the broader trend in VQA research that aims to improve model interpretability through techniques like Grad-CAM.[9] Understanding why a VQA model makes a particular prediction is crucial for building trust in these systems and for identifying potential failure modes or biases.[30]

However, it is important to acknowledge the limitations of relying solely on captions for interpretability.[23] While question-guided captions offer valuable insights, they are still a textual abstraction of the visual scene and might not always capture the full complexity of the visual reasoning involved in answering a question.[23] Therefore, while PNP-VQA offers enhanced interpretability compared to some other approaches, further research into more comprehensive interpretability methods for VQA remains essential.

7. Recent Advancements and Related Work

Since its introduction, the Plug-and-Play VQA (PNP-VQA) framework has garnered attention within the vision and language research community. One significant development has been its integration into the **LAVIS (Language-and-Vision) library** by Salesforce.[14] LAVIS is an open-source deep learning library designed to facilitate research and development in the field of language-vision intelligence.[14] It provides a unified interface for accessing state-of-the-art models and datasets across various multimodal tasks.[36] The inclusion of PNP-VQA in LAVIS, along with the availability of its code and pre-trained models [14], has made the framework more accessible to researchers and practitioners, enabling easier experimentation and further development.[14]

Recent research has also seen the development of systems that build upon or utilize PNP-VQA as a component. For instance, **CodeVQA** is a framework that employs program synthesis to answer visual questions and uses PNP-VQA as a module for answering questions about single images.[4] This demonstrates the versatility of

PNP-VQA's approach and its potential to be integrated into more complex VQA systems to handle specific sub-tasks.[4]

PNP-VQA belongs to a broader category of zero-shot VQA methods that leverage large language models and image captioning.[3] Other notable methods in this space include PICa and Img2LLM.[3] While PICa also converts images to captions for use with a language model, PNP-VQA distinguishes itself by generating multiple question-guided captions and employing a fusion mechanism.[3] Img2LLM further refines this approach by generating question-answer pairs based on the captions to prompt the language model.[42] These methods, including PNP-VQA, represent a significant direction in zero-shot VQA research, aiming to harness the power of pre-trained models without task-specific training.

Given its performance on the OK-VQA dataset, PNP-VQA also contributes to the advancements in **Knowledge-Based VQA (KB-VQA)**.[19] KB-VQA is a challenging subfield that requires models to utilize external knowledge to answer questions about images.[45] Recent trends in KB-VQA involve leveraging large language models not only as knowledge sources but also for complex reasoning.[45] PNP-VQA's ability to perform reasonably well on OK-VQA indicates its potential in this domain, although more specialized KB-VQA methods are continuously being developed.[21]

Finally, the evaluation of VQA models, including PNP-VQA, is an ongoing area of research. There are continuous efforts to develop more robust and human-aligned evaluation metrics, particularly for open-ended answers and in out-of-distribution settings.[34] These advancements in evaluation methodologies will likely provide a more nuanced understanding of the capabilities and limitations of frameworks like PNP-VQA.

8. Method Summary:

Plug-and-Play VQA (PnP-VQA) introduces a zero-shot Visual Question Answering system that combines pretrained modules without additional training. It consists of:

- BLIP for image captioning
- UnifiedQAv2 for question answering
- GradCAM for identifying image regions relevant to the question

9. Task Solved:

Traditional VQA models require end-to-end supervised training, which is inflexible and resource-intensive. PnP-VQA solves this by modularizing the system, enabling adaptability and zero-shot learning.

10. Comparison to Other Methods:

Unlike monolithic VQA models that jointly train vision and language components, PnP-VQA leverages independent, pretrained models. This offers scalability, interpretability, and ease of experimentation.

Performance Conditions:

- Performs well on datasets with structured or synthetic images (e.g., VQAv2 Balanced Real Images) and clear visual-text alignment.
- May underperform on real-world images with complex, unstructured content or when pretrained components have a domain mismatch.

11. Main Claim and Validation

Main Claim:

"Question-guided captioning (via GradCAM-based patch selection) leads to more accurate answers than generic captioning."

Evidence from Paper and Replication:

Initially, we used the Salesforce LAVIS implementation and a subset of 50 examples from the VQAv2 Abstract Scenes validation set, and got an approximate accuracy of 40% for the question-guided pipeline and an approximate accuracy of 36% for generic captioning accuracy. This supports the hypothesis that aligning caption generation with question-relevant regions improves answer quality. However, we stopped using VQAv2 Abstract Scenes because access was denied due to a sudden permission issue on AWS S3.

To continue the replication and improve statistical confidence, we evaluated 10 random 50-image subsets from the VQAv2 Balanced Real Images. The results show that the accuracy between generic captioning and question-guided captioning is very close. Both show an accuracy rate between 30% to 50%.

```
Running subset 1/10
Subset 1 — Guided: 0.380, Generic: 0.400
Running subset 2/10
Subset 2 — Guided: 0.400, Generic: 0.500
Running subset 3/10
Subset 3 — Guided: 0.400, Generic: 0.360
Running subset 4/10
Subset 4 — Guided: 0.360, Generic: 0.400
Running subset 5/10
Subset 5 — Guided: 0.400, Generic: 0.400
Running subset 6/10
Subset 6 — Guided: 0.420, Generic: 0.420
Running subset 7/10
Subset 7 — Guided: 0.380, Generic: 0.340
Running subset 8/10
Subset 8 — Guided: 0.480, Generic: 0.460
Running subset 9/10
Subset 9 — Guided: 0.360, Generic: 0.440
Running subset 10/10
Subset 10 — Guided: 0.340, Generic: 0.340
```

Experimental Setup:

Dataset: 50 diverse samples from the VQAv2 Balanced Real Images dataset

Models Used:

- blip_caption_base_coco (for captioning)
- UnifiedQAv2 (for QA)
- pnp_vqa_base (PnP pipeline)

Baselines:

- Generic captioning + QA
- Direct prediction (PnP baseline)

Metrics: Accuracy(via sklearn.metrics)

Code Availability: All code is available via Salesforce LAVIS GitHub and Colab Notebook

Independent Validation Feasibility:

Yes, the claim is independently verifiable. Both the code and the dataset are publicly available, and the experiment can be replicated with minimal setup.

12. Application Demo

What Was Demonstrated:

We demonstrated three inference pipelines on selected image-question pairs:

- Question-Guided Captioning: Uses GradCAM to select question-relevant patches for captioning.
- Generic Captioning: Uses the full image for captioning.
- Direct Answering: Uses PnP's default question-to-answer module.

Output Interpretation:

- The QA model performs slightly better when the caption includes details relevant to the question.
- GradCAM focuses on meaningful regions, improving context in the generated captions.
- The notebook outputs clearly show visual attention overlays, captions, and predicted answers.

13. Attribution

Paper: Plug-and-Play VQA (EMNLP 2022)

Codebase: A fork version of Salesforce LAVIS

Dataset: VQAv2 Balanced Real Images

Notebook: Colab Link

14. Critical Reflection

What We Learned:

- The power of modular deep learning systems in zero-shot settings.
- How visual attention (GradCAM) can actively influence downstream tasks rather than just serve as a post hoc explanation.

Challenges:

- Installing the correct dependencies for the LAVIS pipeline in Colab.
- Understanding and integrating different APIs (LAVIS, HuggingFace, Sklearn) for smooth model orchestration.

Suggestions for Improvement:

- Use larger and more diverse datasets to improve generalization.
- Explore alternatives to GradCAM for patch selection.
- Fine-tune components in few-shot scenarios for further performance gains.

15. Conclusion and Future Directions

In conclusion, the PNP-VQA framework represents a significant advancement in the field of Visual Question Answering, particularly within the challenging zero-shot learning paradigm. Its key strengths lie in its ability to effectively leverage the knowledge and capabilities of large pre-trained models without requiring any additional task-specific training. The modular architecture, coupled with the innovative use of question-guided captioning as an intermediate representation, allows PNP-VQA to achieve state-of-the-art performance on several benchmark datasets while also exhibiting remarkable parameter efficiency.

Despite its successes, PNP-VQA also faces certain limitations. Its performance can be affected by the quality of the underlying pre-trained models, and it might struggle with certain types of complex reasoning. The reliance on caption generation can also lead to a loss of some visual information. Nevertheless, the framework's interpretability, enhanced by the question-guided captions and the potential use of GradCAM, offers advantages over many end-to-end trained VQA models.

The significance of PNP-VQA lies in its pioneering approach to zero-shot VQA, demonstrating the potential of effectively conjoining existing pre-trained models for

complex multimodal tasks. This is particularly impactful in applications where labeled VQA data is scarce or expensive to obtain, such as in specialized domains like medical imaging or in rapidly evolving environments.[4]

Future research directions for PNP-VQA could explore the integration of more advanced pre-trained vision-language models for improved caption generation, investigate novel strategies for generating and fusing question-guided captions to capture more nuanced visual information, and explore methods to address limitations in complex reasoning, possibly through integration with specialized reasoning modules. Furthermore, evaluating the robustness of PNP-VQA to various visual and textual perturbations and applying it to more specialized domains like medical VQA could be valuable avenues for future work.[3,4] Ultimately, the modular and training-free nature of PNP-VQA underscores the potential of such approaches in advancing the field of multimodal artificial intelligence, paving the way for more adaptable and efficient systems that can understand and reason about both visual and textual information in a wide range of scenarios.

Works Cited:

1. Plug-and-Play VQA: Zero-shot VQA by Conjoining Large Pretrained ..., accessed May 12, 2025, https://aclanthology.org/2022.findings-emnlp.67/
2. [2210.08773] Plug-and-Play VQA: Zero-shot VQA by Conjoining Large Pretrained Models with Zero Training - arXiv, accessed May 12, 2025, https://arxiv.org/abs/2210.08773
3. Plug-and-Play VQA: Zero-shot VQA by Conjoining Large Pretrained Models with Zero Training - ACL Anthology, accessed May 12, 2025, https://aclanthology.org/2022.findings-emnlp.67.pdf
4. Modular visual question answering via code generation - Google Research, accessed May 12, 2025, https://research.google/blog/modular-visual-question-answering-via-code-generation/
5. ar5iv.labs.arxiv.org, accessed May 12, 2025, https://ar5iv.labs.arxiv.org/html/2210.08773#:~:text=In%20contrast%20to%20most%20existing,additional%20training%20of%20the%20PLMs.
6. [2210.08773] Plug-and-Play VQA: Zero-shot VQA by Conjoining Large Pretrained Models with Zero Training - ar5iv, accessed May 12, 2025, https://ar5iv.labs.arxiv.org/html/2210.08773
7. Achieving Zero-Shot SOTA in VQA with Plug-and-Play Framework - Toolify.ai, accessed May 12, 2025, https://www.toolify.ai/ai-news/achieving-zeroshot-sota-in-vqa-with-plugandplay-framework-1938826

8. Modularized zero-shot VQA with pre-trained models - InK@SMU.edu.sg, accessed May 12, 2025, https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=9310&context=sis_research
9. Modularized Zero-shot VQA with Pre-trained Models - arXiv, accessed May 12, 2025, https://arxiv.org/html/2305.17369v2
10. arXiv:2305.17369v2 [cs.CV] 24 Jan 2024, accessed May 12, 2025, https://arxiv.org/pdf/2305.17369
11. ViCrop: Perceiving Small Visual Details in Zero-shot Visual Question Answering with Multimodal Large Language Models - arXiv, accessed May 12, 2025, https://arxiv.org/html/2310.16033v2
12. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization | Facebook AI Research - Meta AI, accessed May 12, 2025, https://ai.meta.com/research/publications/grad-cam-visual-explanations-from-deep-networks-via-gradient-based-localization/
13. Modular Visual Question Answering via Code Generation - ACL Anthology, accessed May 12, 2025, https://aclanthology.org/2023.acl-short.65.pdf
14. Salesforce Open-Sources Language-Vision AI Toolkit LAVIS - InfoQ, accessed May 12, 2025, https://www.infoq.com/news/2022/11/salesforce-lavis-ai/
15. VQA v2 test-dev Benchmark (Visual Question Answering (VQA)) | Papers With Code, accessed May 12, 2025, https://paperswithcode.com/sota/visual-question-answering-on-vqa-v2-test-dev
16. VQA v2 val Benchmark (Visual Question Answering (VQA)) - Papers With Code, accessed May 12, 2025, https://paperswithcode.com/sota/visual-question-answering-on-vqa-v2-val
17. Plug-and-Play VQA: Zero-shot VQA by Conjoining Large Pretrained ..., accessed May 12, 2025, https://paperswithcode.com/paper/plug-and-play-vqa-zero-shot-vqa-by-conjoining
18. salesforce/LAVIS: LAVIS - A One-stop Library for Language-Vision Intelligence - GitHub, accessed May 12, 2025, https://github.com/salesforce/LAVIS
19. aclanthology.org, accessed May 12, 2025, https://aclanthology.org/2024.findings-eacl.36.pdf
20. Knowledge generation for zero-shot knowledge-based VQA - InK@SMU.edu.sg, accessed May 12, 2025, https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=9729&context=sis_research
21. OK-VQA Benchmark (Visual Question Answering (VQA)) | Papers With Code, accessed May 12, 2025, https://paperswithcode.com/sota/visual-question-answering-on-ok-vqa
22. Modularized Zero-shot VQA with Pre-trained Models - ACL Anthology, accessed May 12, 2025, https://aclanthology.org/2023.findings-acl.5.pdf
23. A Simple Baseline for Knowledge-Based Visual Question Answering - ACL Anthology, accessed May 12, 2025, https://aclanthology.org/2023.emnlp-main.919.pdf

24. Plug-and-Play VQA: Zero-shot VQA by Conjoining Large Pretrained Models with Zero Training - Papers With Code, accessed May 12, 2025, https://paperswithcode.com/paper/plug-and-play-vqa-zero-shot-vqa-by-conjoining/review/
25. Overcoming the Limitations of Learning-Based VQA for Counting Questions with Zero-Shot Learning | Request PDF - ResearchGate, accessed May 12, 2025, https://www.researchgate.net/publication/382229379_Overcoming_the_Limitations_of_Learning-Based_VQA_for_Counting_Questions_with_Zero-Shot_Learning
26. Bidirectional Contrastive Split Learning for Visual Question Answering - AAAI Publications, accessed May 12, 2025, https://ojs.aaai.org/index.php/AAAI/article/view/30158/32054
27. Visual Question Answering Models for Zero-Shot Pedestrian Attribute Recognition: A Comparative Study - ResearchGate, accessed May 12, 2025, https://www.researchgate.net/publication/381803506_Visual_Question_Answering_Models_for_Zero-Shot_Pedestrian_Attribute_Recognition_A_Comparative_Study
28. Visual Question Answering using Large Language Models - Munich Data Science Institute, accessed May 12, 2025, https://www.mdsi.tum.de/fileadmin/w00cet/di-lab/pdf/PreciBake_WS2023_FinalReport.pdf
29. Mixture of Rationale: Multi-Modal Reasoning Mixture for Visual Question Answering - arXiv, accessed May 12, 2025, https://arxiv.org/html/2406.01402v1
30. Grad-CAMO: Learning Interpretable Single-Cell Morphological Profiles from 3D Cell Painting Images - CVF Open Access, accessed May 12, 2025, https://openaccess.thecvf.com/content/CVPR2024W/CVMI/papers/Gopalakrishnan_Grad-CAMO_Learning_Interpretable_Single-Cell_Morphological_Profiles_from_3D_Cell_Painting_CVPRW_2024_paper.pdf
31. The employed VQA architecture. | Download Scientific Diagram - ResearchGate, accessed May 12, 2025, https://www.researchgate.net/figure/The-employed-VQA-architecture_fig2_388459622
32. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization - AI Robolab, accessed May 12, 2025, https://airobolab.uni.lu/wp-content/uploads/sites/76/2021/01/Topic-8-Grad-CAM-Presentation.pdf
33. Visual Question Answering Model - Captum · Model Interpretability for PyTorch, accessed May 12, 2025, https://captum.ai/tutorials/Multimodal_VQA_Interpret
34. Visual Robustness Benchmark for Visual Question Answering (VQA) - ResearchGate, accessed May 12, 2025, https://www.researchgate.net/publication/382065234_Visual_Robustness_Benchmark_for_Visual_Question_Answering_VQA
35. A Simple Baseline for Knowledge-Based Visual Question Answering - OpenReview, accessed May 12, 2025, https://openreview.net/pdf?id=3XDDWCu8CF
36. anthonytmh/lavis-pnpvqa: LAVIS - A One-stop Library for Language-Vision

Intelligence, accessed May 12, 2025, https://github.com/anthonytmh/lavis-pnpvqa
37. Welcome to LAVIS's documentation! — LAVIS documentation, accessed May 12, 2025, https://opensource.salesforce.com/LAVIS/latest/index.html
38. LAVIS/README.md at main · salesforce/LAVIS · GitHub, accessed May 12, 2025, https://github.com/salesforce/LAVIS/blob/main/README.md?plain=1
39. LAVIS: A One-stop Library for Language-Vision Intelligence - ACL Anthology, accessed May 12, 2025, https://aclanthology.org/2023.acl-demo.3.pdf
40. LAVIS/README.md at main · salesforce/LAVIS · GitHub, accessed May 12, 2025, https://github.com/salesforce/LAVIS/blob/main/README.md
41. accessed December 31, 1969, https://github.com/salesforce/LAVIS/blob/main/projects/pnp-vqa/README.md
42. Diversify, Rationalize, and Combine: Ensembling Multiple QA Strategies for Zero-shot Knowledge-based VQA - ACL Anthology, accessed May 12, 2025, https://aclanthology.org/2024.findings-emnlp.84.pdf
43. arXiv:2406.01402v1 [cs.CV] 3 Jun 2024, accessed May 12, 2025, https://arxiv.org/pdf/2406.01402?
44. Learning to Ask Denotative and Connotative Questions for Knowledge-based VQA - ACL Anthology, accessed May 12, 2025, https://aclanthology.org/2024.findings-emnlp.487.pdf
45. A Comprehensive Survey of Knowledge-Based Vision Question Answering Systems: The Lifecycle of Knowledge in Visual Reasoning Task - arXiv, accessed May 12, 2025, https://arxiv.org/html/2504.17547v1
46. Ask Me Anything: Free-Form Visual Question Answering Based on Knowledge from External Sources - ResearchGate, accessed May 12, 2025, https://www.researchgate.net/publication/311610816_Ask_Me_Anything_Free-Form_Visual_Question_Answering_Based_on_Knowledge_from_External_Sources
47. (PDF) A Comprehensive Survey of Knowledge-Based Vision Question Answering Systems: The Lifecycle of Knowledge in Visual Reasoning Task - ResearchGate, accessed May 12, 2025, https://www.researchgate.net/publication/391120623_A_Comprehensive_Survey_of_Knowledge-Based_Vision_Question_Answering_Systems_The_Lifecycle_of_Knowledge_in_Visual_Reasoning_Task
48. Visual Robustness Benchmark for Visual Question Answering (VQA) - CVF Open Access, accessed May 12, 2025, https://openaccess.thecvf.com/content/WACV2025/papers/Ishmam_Visual_Robustness_Benchmark_for_Visual_Question_Answering_VQA_WACV_2025_paper.pdf
49. Improving Automatic VQA Evaluation Using Large Language Models - arXiv, accessed May 12, 2025, https://arxiv.org/html/2310.02567v2
50. [2301.12032] BinaryVQA: A Versatile Test Set to Evaluate the Out-of-Distribution Generalization of VQA Models - arXiv, accessed May 12, 2025, https://arxiv.org/abs/2301.12032
51. Reliable Visual Question Answering: Abstain Rather Than Answer Incorrectly - UC Berkeley EECS, accessed May 12, 2025, https://www2.eecs.berkeley.edu/Pubs/TechRpts/2022/EECS-2022-137.pdf
52. Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual

Question Answering - CVF Open Access, accessed May 12, 2025, https://openaccess.thecvf.com/content_cvpr_2017/papers/Goyal_Making_the_v_CVPR_2017_paper.pdf
53. Evaluating State-of-the-Art Visual Question Answering Models Ability to Answer Complex Counting Questions | Journal of Student Research, accessed May 12, 2025, https://www.jsr.org/hs/index.php/path/article/view/2446
54. VQA Challenge 2021 - VQA: Visual Question Answering, accessed May 12, 2025, https://visualqa.org/challenge.html
55. Vision–Language Model for Visual Question Answering in Medical Imagery - PMC, accessed May 12, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10045796/
56. Understanding Visual Question Answering (VQA) in 2025 - viso.ai, accessed May 12, 2025, https://viso.ai/deep-learning/understanding-visual-question-answering-vqa/
57. Large Models in Dialogue for Active Perception and Anomaly Detection - arXiv, accessed May 12, 2025, https://arxiv.org/html/2501.16300v1
58. Comparison with the state-of-the-art methods on the VQA-v2 dataset. - ResearchGate, accessed May 12, 2025, https://www.researchgate.net/figure/Comparison-with-the-state-of-the-art-methods-on-the-VQA-v2-dataset_tbl2_343981315
59. VE-Bench: Subjective-Aligned Benchmark Suite for Text-Driven Video Editing Quality Assessment, accessed May 12, 2025, https://ojs.aaai.org/index.php/AAAI/article/view/32763/34918