Bridging the Selectivity Gap in Zero-Shot Visual Question Answering via Intrinsic Answer Verification

1st Xuan Tang
School of Computer Engineering and Science
Shanghai University
Shanghai, China
xuantang2023@shu.edu.cn

2nd Liyan Ma*
School of Computer Engineering and Science
Shanghai University
Shanghai, China
liyanma@shu.edu.cn

3rd Xiangfeng Luo
School of Computer Engineering and Science
Shanghai University
Shanghai, China
luoxf@shu.edu.cn

***Abstract*—Zero-shot visual question answering (VQA) has recently benefited from large language models that utilize generated captions as intermediate reasoning contexts. We observe that while high Answer Hit Rates (AHR) in captions are a prerequisite for visually grounded reasoning, they do not guarantee reliable final predictions. Specifically, our approach integrates two key components to address this limitation. First, we introduce Question-Guided Context Refinement (QGCR) to generate semantically aligned captions, designed to enrich the extraction of question-critical visual details. Second, to ensure the correct answer is selected, we design an intrinsic answer verification mechanism based on the internal representation consistency of a frozen vision-language encoder. Extensive experiments on GQA and OK-VQA demonstrate that our framework consistently improves final accuracy, showing the effectiveness of intrinsic verification in strictly training-free reasoning scenarios.**

***Index Terms*—Visual Question Answering, Zero-Shot Learning, Vision-Language Models, Answer Verification, Training-Free**

I. Introduction

Visual Question Answering (VQA) stands as a cornerstone for assessing multimodal reasoning. While state-of-the-art models like [1]–[5] achieve impressive performance. However, these advances often depend on extensive pretraining or task-specific fine-tuning, resulting in considerable computational and data costs.

To alleviate the high cost of training, increasing attention has been directed toward zero-shot VQA, where frozen pretrained models are repurposed to answer visual questions without gradient updates. Existing zero-shot VQA methods can be broadly categorized according to how additional information is introduced and controlled during inference: (1) direct multimodal inference depending on the implicit knowledge of pretrained vision–language models [2], [4], [6] ; (2) retrieval- or knowledge-augmented approaches that incorporate external examples or factual evidence at inference time [7]–[9]; and (3) structured, modular reasoning pipelines that explicitly decompose VQA into interpretable stages to improve controllability and reliability without task-specific training [10], [11] . Approaches such as PNP-VQA [10] demonstrate that modular pipelines are highly interpretable and flexible. They treat visual grounding and reasoning as separate steps. As a result, it can seamlessly integrate various pre-trained models without the need for further training.

Despite these advantages, the reliability of modular zero-shot VQA systems remains constrained by the quality of the intermediate context. Through empirical observation, we identify a critical disconnect between information retrieval and final reasoning, which manifests in two aspects. We observe that the answer hit rate (AHR) is positively correlated with the final accuracy in zero-shot VQA settings. Since these models rely entirely on explicitly provided textual context, correct reasoning is strongly correlated with whether the generated captions contain answer-relevant information. However, AHR alone is not sufficient to offer a feasible predictions. More critically, we identify a pronounced selectivity gap in modular zero-shot VQA: even when the correct answer is explicitly present in the generated captions, frozen language models frequently hallucinate or select plausible distractors. This indicates that simply "seeing" the answer is insufficient, the model lacks an intrinsic mechanism to verify and distinguish the correct answer from noise.

Prior works [11]–[15] either retrain or scale up models for better captions, incurring high costs, or remain training-free [11], [13], [14] yet rely on decoder probabilities or heuristic scores prone to hallucination. In contrast, we shift answer verification to the intrinsic mechanisms of a frozen vision-language encoder: correct answers induce a distinctive representation consistency that hallucinations lack. Exploiting this intrinsic signal without extra parameters delivers robust, zero-shot selection.

Based on the above, we make the following contributions:

- We propose a novel strategy to verify answers by checking the representation consistency within frozen encoders. This effectively mitigates hallucinations without the need for external feedback or training.
- We propose a question-guided context refinement module. Specifically, we design distinct prompts for different

question types and apply targeted feature weighting. This allows the generated captions to include specific details required for reasoning.
- We validate our framework through comprehensive evaluations on GQA and OK-VQA. The results show that our training-free approach achieves superior performance compared to existing modular baselines, yielding absolute accuracy gains of 0.7% on OK-VQA and 1.4% on GQA.

Upon acceptance, we will release the source code publicly. The link will be provided in the camera-ready version.

II. Related Work

A. Zero-Shot Visual Question Answering

With the rapid advancement of large language models (LLMs), leveraging their strong reasoning capabilities for zero-shot visual question answering (VQA) has become a prevailing research trend. End-to-end multimodal models [2]–[4], [16], [17], empowered by large-scale vision–language pretraining, have demonstrated impressive performance on VQA benchmarks. However, these methods rely heavily on large amounts of data and require extensive, resource-intensive training, resulting in extremely high costs. They are not suitable when resources are limited. Consequently, modular frameworks, which compose off-the-shelf multimodal models with frozen LLMs, have emerged as a resource-efficient alternative.

To alleviate these limitations, recent studies [10], [12], [15], [18]–[20] have explored modular zero-shot VQA frameworks. These approaches decompose visual perception,language generation,and reasoning into independent components,then chain them together with frozen pretrained models.With offering flexibility and training efficiency,they commonly suffer from two fundamental gaps: (i) a modality disconnect between vision and language, and (ii) a task disconnect between language modeling and question answering. Among such frameworks, Plug-and-Play VQA [10] stands out as a representative instantiation. It systematically couples visual-description generation with text-based question answering, providing a strong training-free baseline upon which later works investigate better context construction and reasoning reliability.

B. Intermediate Context Generation for VQA

Leveraging natural language as an intermediate representation to bridge the modality gap has become a prevailing paradigm for applying LLMs to VQA tasks. Some studies [12], [18] pioneered this approach by converting images into captions to prompt frozen LLMs. Building on this, Prophet [15] enhanced the prompting strategy by incorporating the question, captions, and heuristic answer candidates into a structured format to guide the reasoning process. However, these methods typically rely on generic image captions, which often fail to capture the specific visual details required to answer a given question. To address this limitation, PromptCap [19] argued that captions should be question-specific rather than generic, proposing a model trained to generate descriptions containing answer-related details. Similarly, PNP-VQA [10] utilized network interpretation to synthesize question-relevant captions for downstream reasoning. More recently, Diff-ZSVQA [11] leveraged the generative power of Versatile Diffusion [21] to synthesize exhaustive and fine-grained descriptions, fundamentally aiming to improve performance by maximizing the information density of the intermediate context. Despite its effectiveness, this approach incurs significant computational overhead and relies on heavy backbone models. In contrast, our work seeks to enhance caption quality and question alignment without the need for larger parameter scales or excessive computational costs.

C. Answer Verification in VQA

The modular VQA pipeline employs LLMs during inference to generate answers based on image captions. However, it still suffers from cases where the caption contains sufficient information, yet the answer is incorrect due to model biases or hallucinated reasoning. RankVQA [22] refines the relative ranking of candidate answers, while SAR performs entailment checking and re-ranks the answers according to their entailment scores, ultimately selecting the one with the highest entailment as the final result. A recent method introduces the DAVR [23], which assigns reliability scores to answers through two complementary evaluation paths, aiming to reduce hallucinations. However, it specifically trains a verification model, which under a different zero-shot setting. Inspired by interpretability studies [24]–[28] that associate hallucinations with internal activation patterns, our work differs from prior verification approaches by assessing representation consistency within a frozen encoder, without introducing additional supervision.

III. Methods

We adopt PNP-VQA [10] as our foundational architecture and enhance it with two complementary modules to address the aforementioned challenges.

A. Question-Guided Context Refinement

To ensure the visual input aligns with the question's semantic intent, we propose a fine-grained, question-guided patch sampling strategy. First, we utilize Grad-CAM to compute the gradient-weighted attention maps between the question tokens and image patches. Crucially, we recognize that visual features are hierarchical: lower layers of the ViT (e.g., layers 0–4) capture low-level details like texture and color, middle layers (5–8) focus on structure and shape, while upper layers (9–11) encode high-level semantic concepts. Based on this insight, we introduce a training-free feature reweighting mechanism. Upon receiving a question, we employ a keyword matching rule to identify its semantic focus (e.g., questions containing "color" map to the low-level texture group). For questions that do not trigger specific keyword rules, we default to a uniform weighting strategy across all layer groups to ensure comprehensive feature coverage. We then assign higher weights to the Grad-CAM maps derived from the corresponding layers while suppressing irrelevant noise from mismatched layers. Finally,

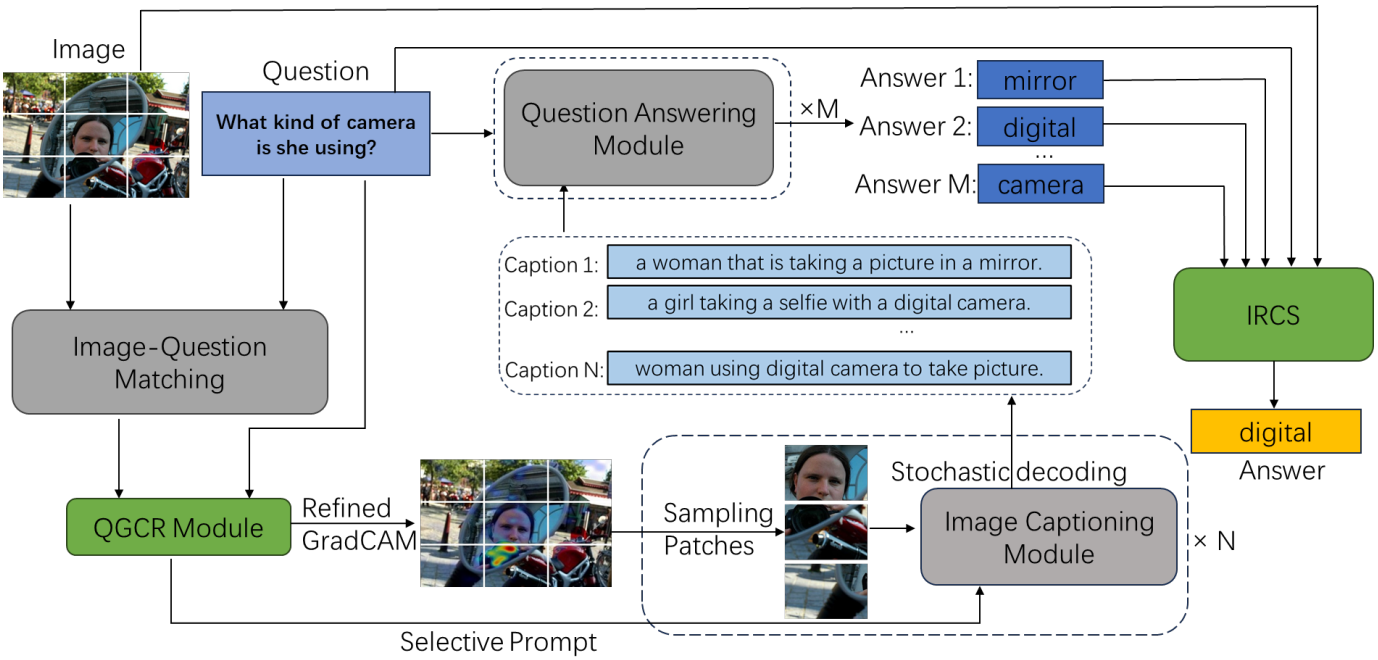

Fig. 1. Overview of the proposed training-free zero-shot VQA framework. The framework extends the PNP-VQA pipeline with (1) a Question-Guided Context Refinement (QGCR) module for question-conditioned visual grounding and caption construction, and (2) an Internal Representation Confidence Scoring (IRCS) module for intrinsic answer verification using a frozen QA-ViT encoder.

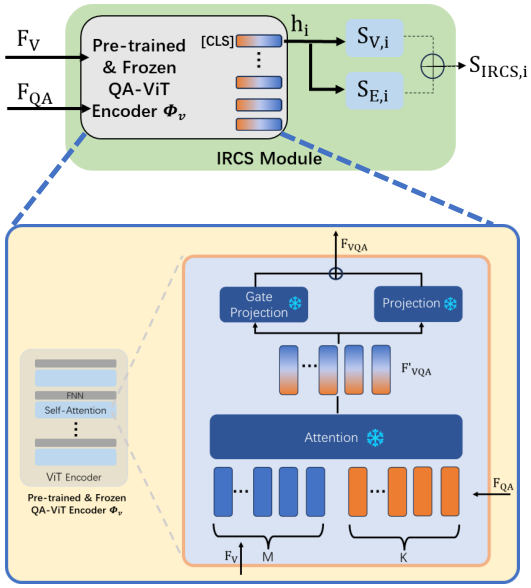

Fig. 2. Architecture of the proposed Internal Representation Confidence Scoring (IRCS) module. Given an image, a question, and a candidate answer, feeding them into a frozen QA-ViT encoder yields the instruction-conditioned global representation from the final-layer [CLS] token, which is then computed to produce an Internal Representation Confidence score; the candidate with the highest IRC score is selected as the final answer.

using these re-weighted relevance scores as a probability distribution, we perform probabilistic sampling to select K' patches. These patches serve as the refined visual input for the subsequent captioning module.

While selected patches provide relevant visual cues, the caption generation model requires explicit guidance to describe them accurately. Previous work using static, generic prompts (e.g., "a picture of") often generate correct but irrelevant captions that diverge from the question's intent. Conversely, directly injecting the question into the prompt (e.g., "a photo showing [question]")—while seemingly direct—proves fragile, leading to performance degradation on certain question types. To address this fragility, we propose a lightweight, keyword-matching-based prompt switching strategy. For question types identified as sensitive to prompt perturbation (e.g., "Who", "What kind"), we utilize the prompt ("a picture of"). Whereas, for object- or attribute-centric questions (e.g., "What color", "What is") where the baseline often lacks focus, we employ the question-injected prompt ("a photo showing [question]") to enforce stricter grounding specificity.

B. Internal Representation Confidence Scoring

We find that even when QGCR ensures the caption contains relevant details is presented in the generated caption, the downstream language model may still fail to select it. This reveals a selectivity gap in modular zero-shot VQA. In response, we seek to improve accuracy through a discriminative approach at a low cost. In particular, we leverage the architectural property of the QA-ViT encoder [29], whose attention layers are modified to accept textual instructions as an additional input. These instruction embeddings are injected directly into the Transformer blocks and participate as key and value states in cross-attention.

Given an image I, a question Q, and a candidate answer A_i, we formulate a declarative instruction $T_i = Question :$

TABLE I

PERFORMANCE COMPARISON WITH PRIOR ZERO-SHOT VQA METHODS ON OK-VQA AND GQA BENCHMARKS. METHODS ARE CATEGORIZED BY THEIR VISION-LANGUAGE ALIGNMENT STRATEGY. **BOLD** INDICATES THE BEST PERFORMANCE; UNDERLINE DENOTES THE SECOND BEST. NOTE THAT DIFF-ZSVQA RELIES ON A 175B LLM, WHEREAS OUR METHOD UTILIZES A LIGHTWEIGHT 223M MODEL.

Method	Language			Vision			OK-VQA	GQA
	Model	#Params	VL-aware	Model	#Params	VL-aware	Test	Test-dev
Pretrained models conjoined by end-to-end VL training.								
VL-T5$_{\text{no-vqa}}$	T5	224M	✓	Faster R-CNN	64M	✗	5.8	6.3
FewVLM$_{\text{base}}$	T5	224M	✓	Faster R-CNN	64M	✗	11.6	27.0
FewVLM$_{\text{large}}$	T5	740M	✓	Faster R-CNN	64M	✗	16.5	29.3
VLKD$_{\text{ViT-B/16}}$	BART	407M	✓	ViT-B/16	87M	✓	10.5	-
VLKD$_{\text{ViT-L/14}}$	BART	408M	✓	ViT-L/14	305M	✓	13.3	-
Flamingo$_{\text{3B}}$	Chinchilla-like	2.6B	✓	NFNet-F6	629M	✓	41.2	-
Flamingo$_{\text{9B}}$	Chinchilla-like	8.7B	✓	NFNet-F6	629M	✓	<u>44.7</u>	-
Flamingo$_{\text{80B}}$	Chinchilla	80B	✓	NFNet-F6	629M	✓	**50.6**	-
Frozen	GPT-like	7B	✗	NF-ResNet-50	40M	✓	5.9	-
Pretrained models conjoined by natural language and zero training.								
PICa	GPT-3	175B	✗	VinVL-Caption	259M	✓	17.7	-
Diff-ZsVQA$_{\text{175B}}$	GPT-3	175B	✗	Versatile Diffusion	≈1B	✓	**45.9**	-
PNP-VQA$_{\text{base}}$(Baseline)	UnifiedQAv2	223M	✗	BLIP-Caption	446M	✓	23.0	34.6
Baseline + QGCR	UnifiedQAv2	223M	✗	BLIP-Caption	446M	✓	23.4	35.1
Baseline + IRCS	UnifiedQAv2	223M	✗	BLIP-Caption	446M	✓	23.3	<u>35.8</u>
Ours (Combine)	UnifiedQAv2	223M	✗	BLIP-Caption	446M	✓	<u>23.7</u>	**36.0**

$QAnswer : A_i$ and feed the triplet (I, T_i) into the frozen QA-ViT encoder. Let h_i denote the final-layer CLS token representation, which summarizes the instruction-conditioned visual features. We analyze the last-layer hidden states of the frozen encoder given different (Image, Question, Answer) triplets. We observe that when the candidate answer matches the visual content, the feature distribution exhibits a significantly higher standard deviation. Concurrently, inverted feature entropy shows high sensitivity to correct answers, indicating a more concentrated and certain feature distribution. Building on this, we leverage this to verify the answers produced by inference. As demonstrated in our experiments, we found that combining these two metrics into a single score yields the same performance to full-model post-verification, and the computational cost is lower. Based on these insights, we propose the IRCS, using a training-free metric that combines feature standard deviation and inverted entropy, to verify answers by assessing the intrinsic consistency of the visual-semantic alignment.

IV. EXPERIMENTS

A. Main result

We conduct experiments under a zero-shot setting on two standard VQA benchmarks: OK-VQA Test [30] and GQA Test-dev [31] . The quantitative results are summarized in Table I. Under similar model parameter setting (approximately 223M), our full model (Ours (combine)) achieves improved accuracy compared to existing zero-shot VQA methods. Compared with the baseline PnP-VQA$_{\text{base}}$, which adopts the same frozen-parameter configuration, our method achieves accuracy improvements of +0.7% on OK-VQA (23.0% $\rightarrow$ 23.7%) and +1.4% on GQA Test-dev (34.6% $\rightarrow$ 36.0%). Although the gains are limited, under strict zero-shot and training-free constraints they reflect a genuine improvement in answer-selection reliability rather than any increase in model capacity. We note that the recently proposed Diff-ZsVQA [11] reports a higher absolute accuracy of 45.9% on OK-VQA. However, this performance is achieved by leveraging a substantially larger GPT-3 (175B) backbone together with a computationally intensive diffusion-based captioning model. Diff-ZsVQA requires substantially more computational resources than our method. The differing performance gains observed on GQA and OK-VQA in Table I are not incidental, but can be attributed to the distinct challenges posed by these two benchmarks. GQA's focus on spatial and compositional reasoning localizes error downstream, so IRCS's zero-parameter, statistics-based diagnosis alone adds +1.2%. OK-VQA contains questions whose answers demand external knowledge that cannot be directly observable from the image. As a result, the performance gains on OK-VQA tend to be smaller than those on GQA. Although reasoning on OK-VQA is primarily constrained by knowledge beyond vision, our method still yields improvements in answer accuracy.

B. Ablation Studies

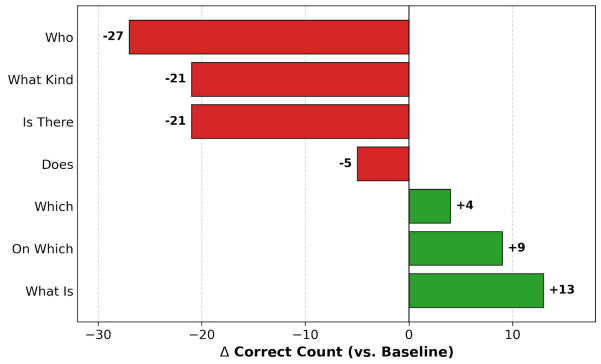

Fig. 3. Prompting sensitivity analysis ($N = 300$ per category). The bars display the net change in correct predictions when switching from the baseline generic prompt ("a picture of") to the question-injected prompt ("a photo showing [question]").

Prompt design offers a low-cost yet effective way to influence caption quality. A straightforward strategy is to inject the question directly into the prompt to enforce question-aware grounding. However, our ablation results, as shown in Fig. 3, reveal that this strategy exhibits strong sensitivity to question types. While question injection significantly improves grounding for object-centric queries such as *What is (+13)* and *On which (+9)*, it leads to substantial performance degradation on abstract or binary questions, including *Who (-27)* and *Is there (-21)*. We attribute this effect to the interaction between question-aware prompting and patch-level visual selection. Although visual patches are sampled stochastically, the sampling distribution is guided by relevance heatmaps. Injecting the question into the prompt may further concentrate attention on question-relevant regions, leading to more consistent patch selection and overly focused descriptions. While such behavior benefits object-centric queries, it can suppress global contextual cues and introduce misleading priors when holistic scene understanding is required. This observation indicates that a single prompt template is not suitable for all types of questions. Motivated by this, we adopt a selective prompting strategy that dynamically chooses prompt templates based on question categories. Based on this flexible design, our method is able to generate image captions that are most conducive to reasoning as much as possible, without adding any additional computational load.

The number of candidate answers determines the size of the verification search space and can therefore significantly affect final accuracy. We first ablate the candidate set size N. As shown in Fig. 4(a) , increasing N from 1 to 5 leads to a steady improvement in accuracy. This suggests that the expanded search space improves answer recall, providing our verification module with a higher chance to identify the correct answer. However, when further increasing N from 5 to 10, we observe a slight performance drop. While a larger candidate set captures more potential answers, it also introduces additional irrelevant candidates, which increases noise during verification. Based on this trade-off, we set the candidate number to

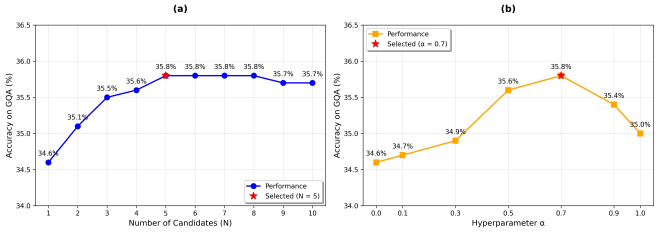

Fig. 4. Ablation studies of the proposed IRCS module on the GQA dataset.(a) Impact of Candidate Number N: Performance improves significantly as the candidate set size expands from $N = 1$ to $N = 5$, peaking at 35.8%.Note that this candidate set size N corresponds to the parameter M illustrated in Fig. 1. (b) Sensitivity of Confidence Weighting α: With N fixed at 5, we evaluate the trade-off between visual feature variance and inverted entropy. The method achieves optimal performance at $\alpha = 0.7$, surpassing both the entropy-only baseline ($\alpha = 0$) and the variance-only setting ($\alpha = 1$).

$N = 5$ in subsequent experiments. Notably, unlike approaches that rely on slow autoregressive decoding for answer verification, our method computes discriminative scores using only a frozen encoder. As a result, even with an enlarged candidate set, our approach maintains low computational overhead.

To further analyze the contribution of each intrinsic signal, we conduct an ablation study on the weighting parameter α, which linearly balances feature standard deviation and inverted feature entropy, as shown in Fig. 4(b). When using feature standard deviation alone ($\alpha = 1$), we see a noticeable improvement in accuracy of approximately 0.4%. This indicates that correct (Image, Question, Answer) triplets induce sharper internal representations, which are driven by visual content rather than language-model priors. The best performance is achieved at $\alpha = 0.7$, which means that visual grounding serves as the dominant signal, while semantic verification remains an indispensable complement. Yet, high visual feature variance alone, although ensuring strong visual grounding, may still admit visually plausible yet semantically incorrect answers. The inverted entropy term suppresses such cases, thereby enhancing semantic consistency.

TABLE II
COMPARISON OF VERIFICATION MECHANISMS.

Architecture	Metric	GQA Acc.	OK-VQA Acc.
Enc-Decoder	Log-Likelihood	35.9%	23.3%
Encoder-Only	**IRC score**	35.8%	23.3%

To further validate our method's capacity to discriminate the correlation within the (Question, Answer, Image) triplet and effectively distinguish visually grounded answers from hallucinations, we compared our method against a direct encoder-decoder architecture for answer verification. Specifically, we employed Flan-T5-Base [32] as the decoder while keeping the encoder identical to ours. In the verification setup, we utilize the T5 model to compute a confidence score for each candidate answer. The process operates in a teacher-forcing manner: the T5 encoder receives the question text and the question-aware visual features (F_{VQ}) as input context, while the candidate answer tokens are supplied to the decoder as the target sequence. We use the length-normalized log-likelihood

of the predicted logits at each token position as the final confidence score. This formulation ensures that the likelihood of the answer is evaluated strictly within the context of the provided question and the re-attended visual features. As shown in Table II, results indicate that our encoder-only IRCS method achieves comparable performance with this encoder-decoder structure on the OK-VQA dataset. Although our method yields a marginally lower accuracy on GQA (-0.1%), this trade-off is accompanied by a substantial reduction in computational cost. While the Encoder-decoder architecture requires computationally intensive autoregressive calculation, our approach operates via a single encoder forward pass. This demonstrates that our method filters hallucinations during reasoning at a lower computational cost.

V. Conclusion

In this work, we identify a selectivity gap in modular zero-shot visual question answering, where correct answers may appear in generated visual descriptions but are still misselected by downstream language models. To address this issue, we study intrinsic answer verification as an independent and training-free component to improve answer selection reliability, while also adopting dynamic prompt design and targeted feature weighting to optimize caption generation. Extensive experiments demonstrate that our method outperforms existing zero-shot VQA methods on mainstream benchmarks under comparable model scales. Despite these gains, our approach remains dependent on the quality of upstream visual descriptions and does not explicitly incorporate external knowledge.

References

[1] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[2] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[3] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 49 250–49 267, 2023.

[4] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[5] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International conference on machine learning*. PMLR, 2021, pp. 5583–5594.

[6] L. Li, J. Peng, H. Chen, C. Gao, and X. Yang, "How to configure good in-context sequence for visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 710–26 720.

[7] L. Gui, B. Wang, Q. Huang, A. G. Hauptmann, Y. Bisk, and J. Gao, "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*, 2022, pp. 956–968.

[8] W. Chen, H. Hu, X. Chen, P. Verga, and W. Cohen, "Murag: Multimodal retrieval-augmented generator for open question answering over images and text," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 5558–5570.

[9] W. Lin and B. Byrne, "Retrieval augmented visual question answering with outside knowledge," *arXiv preprint arXiv:2210.03809*, 2022.

[10] A. M. H. Tiong, J. Li, B. Li, S. Savarese, and S. C. Hoi, "Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training," *arXiv preprint arXiv:2210.08773*, 2022.

[11] Q. Xu, J. Li, Y. Tian, L. Zhou, F. Zhang, and R. Huang, "Diff-zsvqa: Zero-shot visual question answering with frozen large language models using diffusion model," *Expert Systems with Applications*, vol. 275, p. 126951, 2025.

[12] J. Guo, J. Li, D. Li, A. M. H. Tiong, B. Li, D. Tao, and S. Hoi, "From images to textual prompts: Zero-shot visual question answering with frozen large language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 867–10 877.

[13] Ö. Özdemir and E. Akagündüz, "Enhancing visual question answering through question-driven image captions as prompts," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1562–1571, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269137643

[14] Y. Du, J. Li, T. Tang, W. X. Zhao, and J.-R. Wen, "Zero-shot visual question answering with language model feedback," *arXiv preprint arXiv:2305.17006*, 2023.

[15] Z. Yu, X. Ouyang, Z. Shao, M. Wang, and J. Yu, "Prophet: Prompting large language models with complementary answer heuristics for knowledge-based visual question answering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[16] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International Conference on Machine Learning*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256390509

[17] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *ArXiv*, vol. abs/2304.10592, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258291930

[18] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, and L. Wang, "An empirical study of gpt-3 for few-shot knowledge-based vqa," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 3081–3089.

[19] Y. Hu, H. Hua, Z. Yang, W. Shi, N. A. Smith, and J. Luo, "Promptcap: Prompt-guided image captioning for vqa with gpt-3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2963–2975.

[20] D. Khashabi, Y. Kordi, and H. Hajishirzi, "Unifiedqa-v2: Stronger generalization via broader cross-format training," *arXiv preprint arXiv:2202.12359*, 2022.

[21] X. Xu, Z. Wang, E. Zhang, K. Wang, and H. Shi, "Versatile diffusion: Text, images and variations all in one diffusion model," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7720–7731, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253523371

[22] Y. Qiao, Z. Yu, and J. Liu, "Rankvqa: Answer re-ranking for visual question answering," in *2020 IEEE international conference on multimedia and expo (ICME)*. IEEE Computer Society, 2020, pp. 1–6.

[23] X. Wu, Y. Ou, P. Tian, Z. Yang, J. Zhang, P. Li, and L. Gao, "Improving VQA Reliability: A Dual-Assessment Approach with Self-Reflection and Cross-Model Verification," *arXiv e-prints*, p. arXiv:2512.14770, Dec. 2025.

[24] Q. Huang, X. wen Dong, P. Zhang, B. Wang, C. He, J. Wang, D. Lin, W. Zhang, and N. H. Yu, "Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13 418–13 427, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265498818

[25] Y.-S. Chuang, Y. Xie, H. Luo, Y. Kim, J. R. Glass, and P. He, "Dola: Decoding by contrasting layers improves factuality in large language models," *ArXiv*, vol. abs/2309.03883, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261582463

[26] T. Yang, Z. Li, J. Cao, and C. Xu, "Mitigating hallucination in large vision-language models via modular attribution and intervention," in *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*, 2025.

[27] J. Bai, H. Guo, Z. Peng, J. Yang, Z. Li, M. Li, and Z. Tian, "Mitigating hallucinations in large vision-language models by adaptively constraining information flow," *ArXiv*, vol. abs/2502.20750, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:276725283

[28] Y. Sun, Y. Gai, L. Chen, A. Ravichander, Y. Choi, and D. X. Song, "Why and how llms hallucinate: Connecting the dots with subsequence associations," *ArXiv*, vol. abs/2504.12691, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:277857137

[29] R. Ganz, Y. Kittenplon, A. Aberdam, E. Ben Avraham, O. Nuriel, S. Mazor, and R. Litman, "Question aware vision transformer for multimodal reasoning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 13 861–13 871.

[30] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "Ok-vqa: A visual question answering benchmark requiring external knowledge," in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 2019, pp. 3195–3204.

[31] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6700–6709.

[32] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, D. Valter, S. Narang, G. Mishra, A. W. Yu, V. Zhao, Y. Huang, A. M. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, "Scaling instruction-finetuned language models," *ArXiv*, vol. abs/2210.11416, 2022.