Refine and Align: Confidence Calibration through Multi-Agent Interaction in VQA

Ayush Pandey[1], **Jai Bardhan**[1, 2*], **Ishita Jain**[1], **Ramya S Hebbalaguppe**[1], **Rohan Raju Dhanakshirur**[1], **Lovekesh Vig**[1]

[1]TCS Research

ayush.p4@tcs.com, jai.bardhan@cvut.cz, ishita2403@gmail.com, ramya.hebbalaguppe@tcs.com, rohanrd@sit.iitd.ac.in, lovekeshvigin@gmail.com

Abstract

In the context of Visual Question Answering (VQA) and Agentic AI, calibration refers to how closely an AI system's confidence in its answers reflects their actual correctness. This aspect becomes especially important when such systems operate autonomously and must make decisions under visual uncertainty. While modern VQA systems, powered by advanced vision-language models (VLMs), are increasingly used in high-stakes domains like medical diagnostics and autonomous navigation due to their improved accuracy, the reliability of their confidence estimates remains under-examined. Particularly, these systems often produce overconfident responses. To address this, we introduce *AlignVQA*, a debate-based multi-agent framework, in which diverse specialized VLM – each following distinct prompting strategies – generate candidate answers and then engage in two-stage interaction: generalist agents critique, refine and aggregate these proposals. This debate process yields confidence estimates that more accurately reflect the model's true predictive performance. We find that more calibrated specialized agents produce better aligned confidences. Furthermore, we introduce a novel differentiable calibration-aware loss function called *AlignCal* designed to fine-tune the specialized agents by minimizing an upper bound on the calibration error. This objective explicitly improves the fidelity of each agent's confidence estimates. Empirical evaluations across multiple benchmark VQA datasets demonstrate the effectiveness of our approach in significantly reducing calibration discrepancies.

Code — https://github.com/ayushp88/AgenticCalibration

Website — https://refine-align.github.io/

Extended Version — https://arxiv.org/abs/2511.11169

Introduction

Visual Question Answering (VQA) is a foundational task in multimodal artificial intelligence that requires models to jointly process visual content and natural language to generate accurate answers to open-ended questions about images. First introduced to connect vision and language for goal-oriented reasoning (Antol et al. 2015), VQA has evolved into a benchmark for evaluating systems' abilities in compositional reasoning, visual grounding, and language understanding (Agrawal et al. 2016). The task tests a model's capability to extract relevant visual cues, understand complex queries, and synthesize answers—all in a single pipeline.

Agentic architectures for VQA: Recent advancements in VQA have embraced agentic architectures, where multiple interacting agents collaboratively solve complex visual reasoning tasks. For instance, Jiang et al. (Jiang et al. 2024) introduced a zero-shot multi-agent system with specialized experts coordinated adaptively. Hu et al. (Hu et al. 2024) proposed a team of LLM-based agents with tool access, whose outputs are aggregated via voting. Wang et al. (Wang et al. 2023a) designed explainable agents with dedicated roles (Responder, Seeker, Integrator) that operate in a top-down reasoning loop. Similarly, the ARE model (Li et al. 2022) focuses on dynamic action-based reasoning through interaction-aware agents grounded in commonsense. However, these methods often rely on static interaction protocols or fixed agent roles, which limits their adaptability to new types of questions. Additionally, most do not explicitly optimize for confidence calibration, which is critical in high-stakes VQA applications.

Need for Calibration in VQA: Due to its practical relevance, VQA is increasingly being deployed in high-stake real-world domains such as medical diagnosis (Lin et al. 2023; Zhou et al. 2023; Canepa, Singh, and Sowmya 2023), autonomous navigation (Qian et al. 2024; Sima et al. 2025; Marcu et al. 2024; Atakishiyev et al. 2023), and assistive technologies for the visually impaired (Gurari et al. 2018; Chanana et al. 2017). In these settings, it is not only essential for VQA systems to be accurate, but also to be calibrated. A model is said to be calibrated if it's confidence matches the probablity of occurance (Guo et al. 2017). A calibrated model knows when to trust their predictions. Uncalibrated VQAmodels pose serious safety threats. For example, an incorrect but overconfident answer in a medical application can mislead clinicians, while a misjudged scene description in autonomous driving could lead to hazardous decisions. In accessibility tools, a high confident false or misleading answers can reduce user trust posing safety risks. Despite significant improvements in answer accuracy, many leading VQA models are often miscalibrated. i.e. they express high confidence even when their answers are wrong (Groot and Valdenegro-Toro 2024; Zhang et al. 2024). This overconfidence undermines the reliability and interpretability of

*Work done while at TCS Research

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

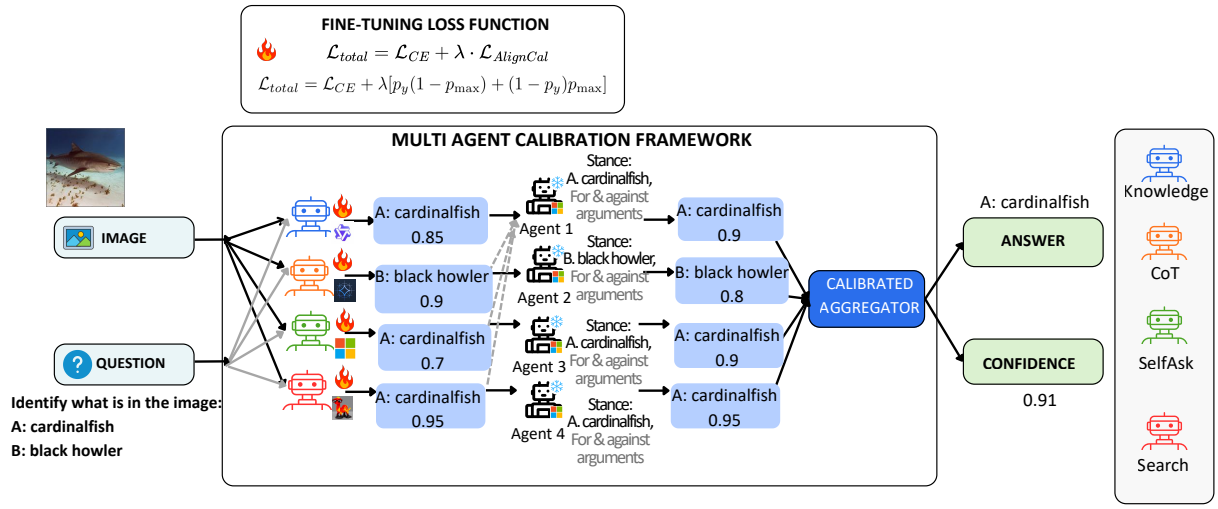


Figure 1: **AlignVQA Multi-Agent Calibration Model.** Given an input image and question, the model first queries a set of specialized agents—*Chain-of-Thought* (Wei et al. 2022), *Search-Augmented*, *SelfAsk* (Press et al. 2022), and *GENREAD Knowledge-based* (Yu et al. 2022) models—each fine-tuned for calibration using our custom proposed loss *AlignCal* . These agents independently produce answer classes (e.g., A: cardinalfish, B: black howler). In the second stage, a group of general agents is instantiated, with each agent probabilistically initialized to a specific answer class based on the distribution of predictions from the specialized agents. These general agents then receive argument-based feedback—comprising *for* and *against* justifications—from all general agents (denoted by dotted grey lines), enabling them to revise both their stance and confidence. The final calibrated prediction is the majority-vote class with associated confidence.

model predictions, especially when the stakes are high.

Measuring Calibration: To evaluate calibration, standard tools such as reliability diagrams and scalar metrics like Expected Calibration Error (ECE) are widely used (Guo et al. 2017; Pakdaman Naeini, Cooper, and Hauskrecht 2015). In this work, we focus on the multiple-choice (MCQ) VQA setting, where the model selects one answer from a predefined set. Calibration in this setting is measured using the standard ECE metric:

$$ECE_{MCQ} := \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right| \quad (1)$$

Here, B_m denotes the set of predictions with confidence values falling into the m-th bin, $\mathrm{acc}(B_m)$ and $\mathrm{conf}(B_m)$ are the average accuracy and confidence in the bin, and n is the total number of samples. A well-calibrated model should produce predictions where the reported confidence matches the actual probability of correctness.

Calibration in SOTA VQA architectures: Several recent works have attempted to improve VQA calibration. Whitehead et al. (Whitehead et al. 2022) proposed a selective answering strategy where the model abstains when it is unsure. Mozaffari et al.'s GLEN framework (Mozaffari, Sapkota, and Yu 2025) introduced a combination of model simplification and focal loss to enhance calibration. IVON by Wieczorek et al. (Wieczorek et al. 2025) leveraged Bayesian variational fine-tuning to capture model uncertainty through posterior weight distributions. While these methods are effective, they come with multiple limitations. Some of them require retraining or fine-tuning the model, while others operate in a single-pass setting without iterative refinement. The potential of multi-agent collaboration for post-hoc calibration remains largely unexplored in the VQA context.

Human-Inspired Calibration via Multi-Agent Debate Humans rarely make decisions in isolation—opinions evolve through discussion, critique, and consensus. This notion of collective wisdom has been explored in recent AI literature through multi-agent debate frameworks that align confidence with justification through interaction (Yang et al. 2024; Oriol et al. 2025; Lin, Hilton, and Evans 2022). Inspired by this, we introduce AlignVQA, a novel calibration method for MCQ VQA. Our approach leverages a structured multi-agent debate, where specialized agents generate initial answers, and generalist agents critique, revise, and update both answers and confidences through a collaborative process. Towards this, we make the following contributions:

1. **Multi-Agent VQA Debate Framework.** We propose a structured post-hoc debate setup in which specialized agents generate initial predictions, and generalist agents iteratively critique and revise these answers and their confidences. The final answer is chosen using a confidence-based aggregation mechanism that maximizes the average confidence across agents. This leads to improved calibration and robustness. Upon applying the proposed agentic framework, the ECE on the VQARad dataset is substantially reduced from its initial value of 0.375 (as reported for Gemma 3 4B) to 0.146. Likewise, the Adaptive Calibration Error (ACE) decreases from

0.207 to 0.133, indicating improved calibration.

2. **Differentiable Calibration-Aware Loss:** We propose a novel differentiable loss function, termed *AlignCal*, which is surrogate to minimize a provable upper bound on the ECE. Unlike conventional loss functions such as cross-entropy or focal loss, which primarily focus on prediction accuracy, *AlignCal* jointly optimizes for both answer correctness and calibration, thereby enhancing the reliability of the specialized agents during training. *AlignCal* leads to a reduction of ECE from 0.232 (focal loss finetuned Gemma 3 4B) to 0.058 on ScienceQA dataset. Incorporating the *AlignCal* finetuned agents in debate framework leads to further reduction of ECE to 0.055 and ACE to 0.110.

Related Works

The common calibration techniques used in classification tasks include: (I.) **Train-time Calibration methods** aim to improve confidence estimates during the training phase by modifying the loss function. These methods generally smooth confidence scores in a sample-agnostic manner–applying regularization uniformly across samples. For example, label smoothing (Szegedy et al. 2016) is a popular train-time calibration method which was originally proposed to improve classifier accuracy by computing the cross-entropy with a weighted sum of the one-hot vector with a uniform distribution. Other works include (Ghosal, Hebbalaguppe, and Manocha 2025; Patra et al. 2023; Hebbalaguppe et al. 2022; Lin et al. 2017; Hebbalaguppe et al. 2024) (II.)**Post-hoc Calibration** Post-hoc calibration methods adjust a fully trained model's confidence scores using a separate hold-out set. For example, (Guo et al. 2017) introduced temperature scaling (TS), which smooths confidence by dividing the logits by a scalar T$>$1. Other notable post-hoc techniques include (Bohdal, Yang, and Hospedales 2021; Islam et al. 2021; Hebbalaguppe et al. 2025).

Multi-Agent Calibration in LLM: Collaborative Calibration (Yang et al. 2024) was introduced as a multi-agent deliberation framework where agents share their predictions, confidence estimates, and the reasoning steps to engage in a simulated group dialogue. Agents iteratively refine responses, leading to improved calibration. This post-hoc ensemble method for LLMs inspires our vision-language adaptation.

Calibration in VQA Utilizing a popular strategy of using consistency among samples to estimate confidence, Eisenschlos et al. (Eisenschlos et al. 2024) introduced a method for improving the reliability of visual question answering (VQA) models. Their calibration approach involves generating multiple answers to a given question and computing the expected pairwise BLEU score, weighted by likelihood of each response under the model's sampling distribution. Then, this expected BLEU score is used to estimate the confidence. The authors applied their approach both directly on images and on gold-standard image captions generated by human annotators. However, a notable improvement in calibration metrics was observed only when using captions, highlighting a key limitation of this method.

Multi-Agent Calibration in Other Visual tasks An ensemble technique for Image Classification introduced by Schulze et al. (Schulze et al. 2025) consists of attaching multiple independently trained classifier "heads" to a shared, frozen backbone. The ensemble methods are aggregated using three methods: simple averaging, majority voting, and the use of metamodelings–where the outputs of all heads serve as inputs to a separate model that produces the final prediction. Among these, metamodeling yielded significantly better performance, whereas averaging and majority voting showed minimal improvements. While this approach is computationally efficient and achieves better calibration metrics, its evaluation is limited to relatively small datasets such as CIFAR-100, raising questions about its scalability to larger and more complex benchmarks. Moreover, since metamodeling was the only aggregation strategy to yield meaningful gains, it suggests that the improvement may stem more from the aggregator itself than from the ensemble structure.A related approach in the domain of object detection is MoCaE: Mixture of Calibrated Experts (Oksuz et al. 2023), which aims to improve the accuracy of confidence estimates through expert ensembling.

Multi-Agent approaches in VQA tasks Wang et al. (Wang et al. 2023b) proposed a multi-agent architecture for VQA that draws inspiration from the human process of top-down reasoning—where individuals leverage prior knowledge and contextual cues to infer new information (e.g., predicting rain from observing cloudy skies). Their framework, referred to as Top-Down Reasoning, consists of three specialized agents: the responder, a vision-language model (VLM) that generates answers to visual questions; the seeker, which formulates relevant follow-up questions based on contextual understanding; and the integrator, which synthesizes insights from both agents to produce the final response. While the paper does not address calibration, its multi-agent framework can enhance answer reliability via improved confidence estimates. To our knowledge, no prior work leverages multi-agent methods for calibration in VQA.

Proposed Methodology

Preliminaries

Vision Language Models (VLMs) in VQA are empirically miscalibrated (see Tab. 1), often exhibiting overconfident incorrect answers. To address this, we propose a two-stage calibration strategy built on agentic debate and refinement: (i) first stage of diverse expert answer generation and semantic clustering, and (ii) a second stage confidence refinement through deliberative generalist agents, followed by calibration aware aggregation. The overall design draws inspiration from recent multi-agent debate and refinement strategies applied to LLMs (Yang et al. 2024), and adapts it a VQA setting with VLMs.

The VQA task can be formulated as learning a function $f : \mathcal{I} \times \mathcal{Q} \rightarrow \mathcal{A}$ that, given an image $i \in \mathcal{I}$ and its associated question $q \in \mathcal{Q}$, predicts an answer $\hat{y} \in \mathcal{A}$ and confidence $p \in [0, 1]$.

Architecture	ScienceQA							VQARad						
	↑Acc.	↑F1	↑Prec.	↑Rec.	↓ACE	↓ECE	↓MCE	↑Acc.	↑F1	↑Prec.	↑Rec.	↓ACE	↓ECE	↓MCE
LLAVA One Vision	67.20%	0.380	0.384	0.405	0.338	0.335	**0.366**	47.00%	0.348	0.570	0.507	0.2312	0.232	**0.286**
Gemma 3 4B	71.00%	0.443	**0.582**	0.372	0.398	0.398	0.464	59.40%	0.591	0.606	0.601	0.208	0.375	0.818
Qwen2.5-VL-3B-Inst.	69.70%	**0.482**	0.467	**0.504**	**0.302**	**0.302**	0.702	**69.00%**	**0.710**	**0.700**	**0.680**	0.294	0.295	0.297
Phi-4-multimodal-Inst.	**76.00%**	0.235	0.230	0.254	0.575	0.574	0.657	58.00%	0.580	0.580	0.579	**0.109**	**0.134**	0.425

Table 1: Comparison of state-of-the-art model performance on the ScienceQA and VQARad datasets. Bold values indicate best performance for each metric within each dataset, underlined values indicate second-best performance.

Agent Ensemble and Stance Generation

Given an input image-question pair (for example, "Identify the type of fish in the picture" (c.f. Fig. 1), we first generate candidate answers from a diverse set of *specalized expert agents*. Each agent is created with a different VLM backbone – Qwen2.5-VL-3B-Instruct (Bai et al. 2025), Llava-Onevision(Li et al. 2024), Gemma 3 4B (Team et al. 2025) and Phi-4-multimodal-instruct (Abouelenin et al. 2025) – and a distinct prompting strategy to encourage diverse reasoning[1]. Specifically, we employ: Chain-of-Thought prompting (Wei et al. 2022) for multi-hop reasoning, Self-Ask prompting (Press et al. 2022) for recursive problem decomposition, Search-style prompting startegy to incorporate external retrieval cues and the GENREAD style prompting (Yu et al. 2022) for structured comprehension.

Each expert agent i (independently) produces an output $v_i = (\hat{y}_i, p_i)$, where $\hat{y}_i$ is the answer string and p_i is its sequence probability. We infer the confidence of the sequence through the geometric mean of probabilities of next-tokens generated. This serves as the initial confidence estimation for a candidate answer of a particular agent. Because different agents may produce semantically equivalent but lexically different answers, we merge the semantically equivalent answers into K *semantically unique stances*,

$$\{s_1, s_2, \ldots, s_K\}, \qquad K \leq N$$

where N is the total number of individual responses across agents and prompting strategies, using GPT-3.5 judge following (Tian et al. 2023).

For each stance s_k, we define its index set as

$$\mathcal{I}_k = \{i \mid \hat{y}_i = s_k\}$$

which denotes all positions i such that the i-th answer $\hat{y}_i$ is equal to the stance s_k. The *frequency* of stance s_k is then,

$$f_k = |\mathcal{I}_k|$$

and the *mean confidence* associated with that stance is computed as

$$\bar{c}_k = \frac{1}{f_k} \sum_{i \in \mathcal{I}_k} c_i$$

where c_i denotes the sequence probability of the i-th answer. Stage 1 thus yields a set of triplets,

$$\{(s_k, f_k, \bar{c}_k)\}_{k=1}^{K},$$

capturing both the diversity of opinion and its strength in terms of support and confidence.

[1]This agentic framework is model agnostic, any set of VLM backbones can be substituted in place of those used here.

Illustrated Failure Case. Fig. 1 in the first stage, three agents each give the answer *cardinal fish* with confidence scores of 0.85, 0.70, and 0.95, producing an average confidence of 0.83 for that stance, while a fourth agent answers *black howler* with confidence 0.90. A majority confidence based system would adopt *black howler* as the consensus after the first stage, despite it being incorrect and lacking group support. Stage 2, is designed to revisit and refine such consensus through deliberation and counter argumentation.

Group Debate with Rationale and Feedback

The second stage introduces a set of **generalist deliberation agents** (no specialized prompting) whose role is to critically examine, defend and revise the candidate stances produced in Stage 1, forming a structured debate ensemble. To maintain the prior group consensus while allowing contrarian exploration, each generalist agent j is assigned a stance s_j by sampling proportionally to the frequencies f_k, i.e., $\Pr(s_j = s_k) \propto f_k$. This maintains a soft bias towards majority supported views while still allowing minority stances to be reconsidered.

Each agent then argues for its stance by exploring diverse reasoning and developing rationales for defending it. Each reasoning path is unique and develops an ensemble of rationales for a particular stance. Agents then provide ratings and feedback to each rationale in terms of logical consistency, factuality, clarity and conciseness. Specifically, Chain-of-Verification style prompting (Dhuliawala et al. 2024) is used to check the factuality by generating underlying premises or assumptions. These premises are then further checked with a search augmented agent to identify unfactual statements in the feedback.

Each general agent then receives a pair of arguments one sampled from the set of supporting arguments and one sampled from one of the opposing sides to form a debate pair. This mirrors two-sided deliberation paradigms in multi-agent reasoning and debate systems. Based on these arguments, each agent produces a final answer that incorporates the provided opposing argument, supporting argument and its previously assigned stance. Thus, the final answer is given by $y'_j = f_j(s_j, \bar{c}_j, a_p, a_n)$, where a_p, a_n are the supporting and opposing arguments with ratings and feedback, and $s_j, \bar{c}_j$ is the initial stance with its associated confidence assigned to agent j. We also record the sequence probability of each agent's final response y'_j, which serves as the refined confidence score $\mathrm{Conf}(y'_j)$.

After collecting the set of refined outputs, we

$\{(y'_j, \mathrm{Conf}(y'_j))\}_{j=1}^{M}$ from M generalist agents, we aggregate it in two steps: First for each stance s, we define the agent index set $\mathcal{I}'_s = \{j \mid y'_j = s\}$ and update $f'_k = |\mathcal{I}'_k|$, and compute the mean refined confidence as

$$\widehat{c}_k = \frac{1}{|\mathcal{I}'_k|} \sum_{j \in \mathcal{I}'_k} \mathrm{Conf}(y'_j).$$

The final answer is selected by majority vote by choosing the stance with the most supporting agents:

$$s^* \;=\; \arg \max_{k \,\in\, \{1,\ldots,K\}} f'_k \tag{2}$$

The final confidence is the mean confidence of the agents supporting the chosen stance. This aggregation yields a better indication of prediction, by weighing different arguments through deliberation. To further improve calibration, we introduce the following loss.

Calibration Aware Finetuning

Our system can benefit from better calibrated VLMs. Therefore, we introduce a novel surrogate loss function that directly minimizes a tight upper bound on miscalibration during training, thereby avoiding the pitfalls of post-hoc fixes. Classical metrics like ECE average confidence–accuracy gaps over broad bins, so they often fail to estimate the reliability of a single test example or a specific subpopulation. Consider a tumor screening classifier that reports confidence 0.9 on all five patients: it is correct on three cases (two tumor positives and one healthy negative) and incorrect on two (one missed tumor and one false positive), so the empirical accuracy is $3/5 = 0.6$. Because all predictions fall into the same confidence bin, the standard ECE is $|0.9 - 0.6| = 0.3$, which reflects only the coarse aggregate gap. In contrast, the Upper Bound Calibration Error (UBCE) (Zhong et al. 2025) averages per-instance absolute gaps: correct cases contribute $1 - 0.9 = 0.1$ each and incorrect ones contribute 0.9 each, yielding UBCE $= (3 \times 0.1 + 2 \times 0.9)/5 = 0.42$. The ECE therefore understates the expected misalignment because it conflates the high-confidence errors with high-confidence correct predictions via binning, whereas UBCE exposes the full error—including the overconfidence on the incorrect tumor diagnosis—by aggregating each individual confidence–correctness gap.

Therefore, we construct a differentiable loss by applying the plug-in principle to the sample estimate of UBCE (an upper bound on calibration metrics). Minimizing this surrogate loss directly drives down ECE and provides gains to MCE.

Our loss function *AlignCal*: Formally, given softmax outputs $\mathbf{p} = (p_1, \ldots, p_K)$ with logits z_i, true label y, top predicted confidence $p_{\max} = \max_i p_i$ and predicted ground truth class probability p_y, we define the soft-calibration loss:

$$\mathcal{L}_{\textit{AlignCal}}\,(p_y, p_{\max}) = p_y(1 - p_{\max}) + (1 - p_y)p_{\max} \tag{3}$$

The full training objective therfore becomes:

$$\mathcal{L}_{\text{tot}} = \mathcal{L}_{\text{FL}} + \lambda \mathcal{L}_{\textit{AlignCal}}\,,$$

where λ is a tuneable hyperparameter and $\mathcal{L}_{\text{FL}}$ is the focal loss (Mukhoti et al. 2020) which is a strictly proper scoring rule that encourages accuracy and also is found to implicitly enhance calibration. Our calibration term is not an ad hoc tweak; it arises naturally as a plug-in surrogate to UBCE:

$$\begin{aligned}\text{UBCE} = \Pr(t=0)\mathbb{E}\left[p_{\max} | t=0\right] \\ + \Pr(t=1)\left(1 - \mathbb{E}\left[p_{\max} | t=1\right]\right),\end{aligned}$$

where $t = \mathbb{I}\{\hat{y} = y\}$ is an indicator function for whether the prediction is correct and $\Pr(\cdot)$ defines the probability. Equivalently, algebraic regrouping gives us,

$$\text{UBCE} = \mathbb{E}\left[t(1 - p_{\max}) + (1-t)p_{\max}\right].$$

This form makes it clear that UBCE is the expected absolute gap between the correctness and confidence – hence a conservative upper bound on the calibration metrics. Applying the law of total expectation (tower rule/law of iterated expectations) and conditioning on an input x, we have,

$$\begin{aligned}\text{UBCE} &= \mathbb{E}_x\left[\mathbb{E}\left[t(1 - p_{\max}) + (1-t)p_{\max} | x\right]\right] \\ &= \mathbb{E}_x\left[\mathbb{E}\left[t|x\right](1 - p_{\max}) + \mathbb{E}\left[(1-t)|x\right]p_{\max}\right].\end{aligned}$$

Here the outer expectation is over inputs x drawn from the data distribution, and the inner expectation is over the randomness in the correctness indicator t given x (i.e., over $y \sim P(y \mid x)$). Since $p_{\max}$ is deterministic conditioned on x, it can be pulled outside the inner expectation.

Let $q(x) := \mathbb{E}\left[t|x\right]$ be the true conditional probability of correctness given an input x. Then,

$$\text{UBCE} = \mathbb{E}_x\left[q(x)(1 - p_{\max}) + (1 - q(x))p_{\max}\right].$$

In practice, $q(x)$ is unknown because it depends on a (possibly deterministic) decision rule, such as an indicator function $\mathbb{I}\{\hat{y} = y\}$ which is non-differentiable and does not allow backpropagation. To obtain a differentiable surrogate, we plug in the model's own soft belief about correctness, namely, instead of taking $\hat{y} = \arg\max p_i$, imagine sampling a label $\hat{y} \sim p\,(\cdot|x)$ from the model's softmax output. Then the indicator function $t = \mathbb{I}\{\hat{y} = y\}$ has $\mathbb{E}\left[t|x\right] = p_y$, i.e., the probability that a randomly drawn label equals the true label is exactly p_y. This makes p_y a natural, smoothed estimator of $q(x)$. This replacement, is an instance of the classical plug-in principle for conditional expectation estimation – which is well studied in statistical learning theory (Grunewalder 2018). Accordingly, we define our soft surrogate loss per example as,

$$\mathcal{L}_{\textit{AlignCal}}\,(p_y, p_{\max}) = p_y(1 - p_{\max}) + (1 - p_y)p_{\max} \tag{4}$$

so that,

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}_{\textit{AlignCal}}\right] &= \mathbb{E}_x\left[p_y(1 - p_{\max}) + (1 - p_y)p_{\max}\right] \\ &\approx \text{UBCE},\end{aligned}$$

with equality in expectation under the assumption that $p_y \approx \mathbb{E}\left[t|x\right]$. This construction gives a smooth, differentiable surrogate to UBCE that admits gradient-based optimization while preserving the structure of the original bound.

From a probabilistic calibration perspective, if the model is well calibrated in the sense that its softmax outputs reflect the true posterior – i.e, $p_y \approx \Pr(y|x)$ – then p_y also approximates $\Pr(\hat{y} = y|x)$, making the plug-in substitution for $q(x)$

Architecture	VQARad Dataset							ScienceQA Dataset						
	↑Acc.	↑F_1	↑Prec.	↑Rec.	↓ACE	↓ECE	↓MCE	↑Acc.	↑F_1	↑Prec.	↑Rec.	↓ACE	↓ECE	↓MCE
Agentic Framework	65.70%	0.540	0.554	0.544	0.144	0.146	0.820	72.80%	0.340	0.346	0.328	0.265	0.270	0.438
Post-Hoc Calibration														
Agentic + TS	65.70%	0.540	0.554	0.544	0.114	0.117	0.765	72.80%	0.340	0.346	0.328	0.255	0.268	0.421
Agentic + DC	65.70%	0.540	0.554	0.554	0.097	**0.041**	**0.113**	72.80%	0.340	0.346	0.328	–	–	–
Train-Time Calib.														
Agentic + FL	**68.50%**	0.571	0.542	0.605	0.116	0.073	0.393	74.40%	0.424	0.480	0.381	0.142	0.180	0.678
Agentic + LS	67.70%	**0.650**	**0.652**	**0.650**	0.175	0.183	0.543	75.20%	0.467	0.532	**0.424**	0.186	0.186	0.916
Proposed Method														
Agentic+*AlignCal* +FL	68.20%	0.548	0.517	0.583	**0.095**	0.098	0.267	**76.10%**	**0.472**	**0.540**	0.418	**0.110**	**0.055**	**0.331**

Table 2: Comprehensive comparison of calibration strategies across VQARad and ScienceQA datasets. Bold values indicate best performance for each metric within each dataset, underlined values indicate second-best performance. The proposed method (Agentic + *AlignCal* + FL) demonstrates superior calibration performance with competitive accuracy across both datasets. It is not possible to perform Dirchelet Calibration in the case of ScienceQA Dataset due to the unavailability of the probabilities of other options.

increasingly accurate. Finally, there is a self-correcting feedback: early in training p_y might poorly estimate $q(x)$, but the calibration loss penalizes discrepancies between $p_{\max}$ and this current belief. Improving those discrepancies tends to make p_y "more honest" about the correctness, which in-turn makes the surrogate tighter. This is analogous to consistency of plug-in classifier – if the estimation of the conditional quantity improves, the overall decision or loss converges to the ideal (Grunewalder 2018).

Gradient Analysis of the proposed loss: To understand the learning dynamics induced by the combined objective, we analyze the gradients. We see that

$$\frac{\partial \mathcal{L}_{AlignCal}}{\partial z_i} = (1-2p_y)p_{\hat{y}}(\delta_{i,\hat{y}}-p_i)+(1-2p_{\hat{y}})p_y(\delta_{i,y}-p_i),$$

where $\hat{y} = \arg\max_j p_j$ is the highest predicted class index, and $\delta_{i,j}$ is the Kronecker delta. This expression never collapses for the interior probabilities and thus provides a constant margin based push. When the model makes an overconfident incorrect guess, the loss acts to reduce the confidence on the erroneous top prediction and boosts the true class; when the model is underconfident, the calibration term works in conjunction with the cross-entropy term (focal loss) to increase p_y above the threshold. We can further study the equilibrium case with respect to the loss to find that for a reasonable choice of λ, the combined loss provides a strict gradient for driving $p_y \to 1$, and for cases where $\hat{y} \neq y$, a gradient for driving $p_{\hat{y}} \to 0$.

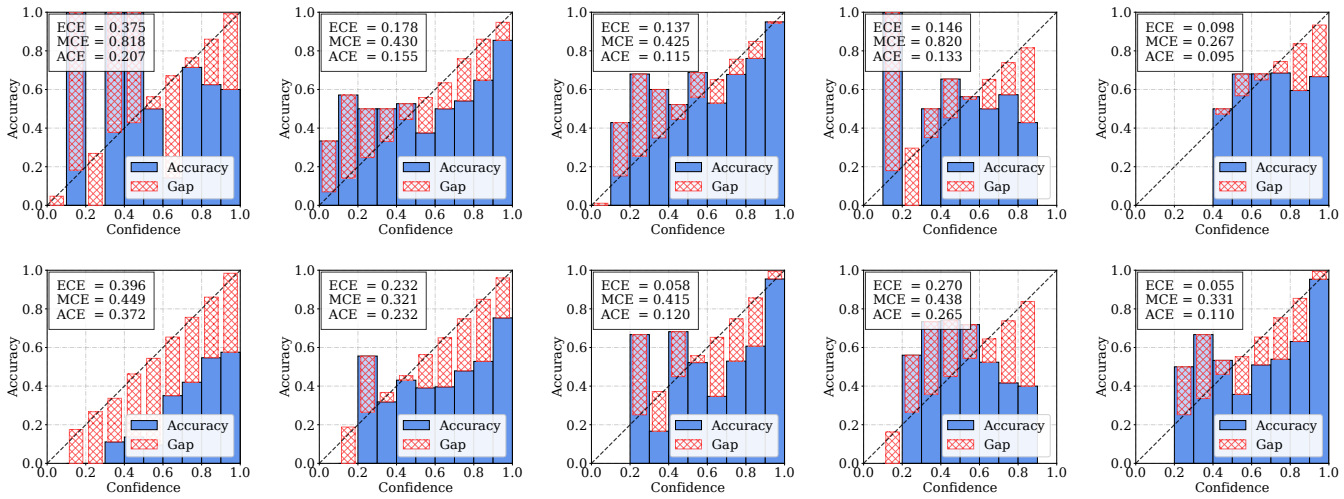
Thus minimizing $\mathcal{L}_{AlignCal}$ directly tightens a provable upper bound on calibration error. Unlike MMCE (Kumar, Sarawagi, and Jain 2018) and other kernelized calibration penalties that require careful choice of kernels and suffer from increased per-update cost, our formulation is closed-form, hyperparameter-light (just λ). In contrast to label smoothing, which bluntly softens all targets and can undermine the sharpness of well-calibrated predictions, our loss targets the disparity between $p_{\max}$ and p_y, dynamically adjusting based on the model's confidence geometry. Focal loss improves calibration only indirectly by reweighting hard examples for better accuracy, while our loss has an explicit probabilistic interpretation as a surrogate to UBCE, ensuring that reducing the training objective systematically reduces ECE by design. Worst-case deviation (MCE) is not guaranteed without extra uniformity constraints since we are minimizing an expectation. Though, we observe practical improvements there as well. Our loss improves over the standard focal loss, as shown in Fig. 2.

Dataset and Evaluation

Datasets and Models: We took 2 publicly available datasets, ScienceQA (Lu et al. 2022) and medical dataset VQARad (Lau et al. 2018). ScienceQA consists of 21,208 multimodal multiple choice questions with diversE science topics and annotations of their answers with corresponding lectures and explanations. VQA-RAD is manually constructed dataset in radiology where answers about images are naturally created and validated by clinicians. VQA-RAD dataset contains 3,515 total visual questions. We only consider Yes/No type of questions from this dataset. The agentic framework consists of four VLM backbones Qwen2.5-VL-3B-Inst. (Bai et al. 2025), Llava-Onevision (Li et al. 2024), Gemma 3 4B (Team et al. 2025), and Phi-4-multimodal-Inst. (Abouelenin et al. 2025). For generalist agent Phi-4-multimodal-instruct, is taken as backbone.
Evaluation: We evaluate calibration with ECE (Guo et al. 2017), ACE, and MCE. To visualize miscalibration, we include reliability diagrams. For task performance, we additionally report Accuracy, F1-score, Precision, and Recall. Maximum Calibration Error (MCE) is the largest absolute difference between predicted confidence and empirical accuracy across all confidence bins.

$$\text{MCE} = \max_{b=1,\dots,B} \left|\text{acc}(B_b) - \text{conf}(B_b)\right| \tag{5}$$

Adaptive Calibration Error (ACE) splits the sorted predictions into bins each containing an equal number of exam-



(a) Gemma 3 4B (b) FL (c) *AlignCal* + FL (d) Agentic Framework (e) Agentic + *AlignCal*

Figure 2: Reliability plots of the datasets, VQARad (top) and ScienceQA (bottom). 2a shows the calibration from base Gemma model. 2b shows plot on FL finetuned Gemma 3 4B model. 2c shows the plot on FL + *AlignCal* finetuned Gemma 3 4B model. 2d shows the plot obtained from Agentic Framework. 2e shows the plot obtained from Agentic framework where agents are finetuned with *AlignCal* + FL.

ples, then computes the mean absolute gap between empirical accuracy and average confidence across those bins.

$$\text{ACE} = \frac{1}{B}\sum_{b=1}^{B}\left|\text{acc}(B_b) - \text{conf}(B_b)\right| \tag{6}$$

Comparison: We compare our multi-agent debate framework (viz. *Agentic Framework*) to the base VLM models, and compare our loss *AlignCal* against other standard calibration techniques Focal Loss (FL) (Mukhoti et al. 2020), Label Smoothing (LS) (Szegedy et al. 2016), and post-hoc methods, including temperature scaling (TS) (Guo et al. 2017) and Dirichlet calibration (DC) (Kull et al. 2019).

Experiment and Results

Agentic Framework We report here the Agentic framework results. From Fig. 2, we can show that agentic VLM debate leads to better calibration and more reliable responses. On ScienceQA dataset ECE decreases from 0.396 to 0.270, while MCE and ACE also show consistent reductions from 0.449 to 0.438 and 0.372 to 0.265, respectively. In the VQARad dataset, ECE decreases from 0.375 to 0.143, ACE also shows consistent reduction 0.207 to 0.144.

***AlignCal* Results.** Fig. 2 reports the calibration improvements achieved by *AlignCal* on two VQA benchmarks. On the VQARad dataset, ECE decreases from 0.178 to 0.137, and ACE decreases from 0.155 to 0.115. Similarly, on the ScienceQA dataset, ECE is reduced from 0.232 to 0.058, while ACE falls from 0.232 to 0.120. We also compared our results with other training calibration methods FL(Lin et al. 2017) and label smoothing (Szegedy et al. 2016). For LS, we use $\alpha = 0.1$ and for FL, we use $\gamma = 2$.

Comparison with Train Time and Post Hoc Calibration techniques. We apply temperature scaling and dirchelet calibration on the results obtained from the Baseline agentic setup (Table 2). ECE improves from 0.1430 to 0.1165, while MCE improves from 0.82 to 0.7634 on applying temperature scaling. Dirchelet calibration improves ECE from 0.1437 to 0.0410 and ACE from 0.1437 to 0.0973. We also compare our results with baseline agents fine tuned with Focal Loss and Label Smoothing.

Debate with Calibrated Agents From Fig. 2, we see that in the Agentic debate framework when the agents are finetuned with *AlignCal* loss, it leads to better calibration. ECE reduces from 0.375 to 0.098, MCE from 0.818 to 0.267 and ACE from 0.207 to 0.095 on VQARad dataset. On ScienceQA dataset ECE reduces from 0.396 to 0.055, ACE reduces from 0.372 to 0.110 and MCE from 0.449 to 0.331. The significant reduction in both across the VQARad and ScienceQA datasets—relative to using fine-tuning or debate in isolation—indicates that the calibrated-agents debate yields substantially more reliable confidence estimates and, consequently, more trustworthy answers.

Supplementary material is included in extended version.

Conclusion

In this work, we propose AlignVQA, a novel method to improve confidence calibration in Visual Question Answering (VQA) using a multi-agent debate framework. AlignVQA tackles overconfident miscalibration in state-of-the-art VQA models, enhancing reliability in high-stakes domains. Our key contribution is the Agentic Debate Framework with a differentiable calibration-aware loss (AlignCal) that effectively reduces miscalibration. Overall, AlignVQA delivers better-calibrated, more reliable answers.

References

Abouelenin, A.; Ashfaq, A.; Atkinson, A.; Awadalla, H.; Bach, N.; Bao, J.; Benhaim, A.; Cai, M.; Chaudhary, V.; Chen, C.; et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Agrawal, A.; Lu, J.; Antol, S.; Mitchell, M.; Zitnick, C. L.; Batra, D.; and Parikh, D. 2016. VQA: Visual Question Answering. arXiv:1505.00468.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Atakishiyev, S.; Salameh, M.; Babiker, H.; and Goebel, R. 2023. Explaining Autonomous Driving Actions with Visual Question Answering. arXiv:2307.10408.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bohdal, O.; Yang, Y.; and Hospedales, T. 2021. Meta-calibration: Learning of model calibration using differentiable expected calibration error. *arXiv preprint arXiv:2106.09613*.

Canepa, L.; Singh, S.; and Sowmya, A. 2023. Visual Question Answering in the Medical Domain. arXiv:2309.11080.

Chanana, P.; Paul, R.; Balakrishnan, M.; and Rao, P. 2017. Assistive technology solutions for aiding travel of pedestrians with visual impairment. *Journal of rehabilitation and assistive technologies engineering*, 4: 2055668317725993.

Dhuliawala, S.; Komeili, M.; Xu, J.; Raileanu, R.; Li, X.; Celikyilmaz, A.; and Weston, J. 2024. Chain-of-Verification Reduces Hallucination in Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 3563–3578. Bangkok, Thailand: Association for Computational Linguistics.

Eisenschlos, J. M.; Maina, H.; Ivetta, G.; and Benotti, L. 2024. Selectively Answering Visual Questions. *arXiv preprint arXiv:2406.00980*.

Ghosal, S. S.; Hebbalaguppe, R.; and Manocha, D. 2025. Better features, better calibration: a simple fix for overconfident networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 231–247. Springer.

Groot, T.; and Valdenegro-Toro, M. 2024. Overconfidence is Key: Verbalized Uncertainty Evaluation in Large Language and Vision-Language Models. arXiv:2405.02917.

Grunewalder, S. 2018. Plug-in Estimators for Conditional Expectations and Probabilities. In Storkey, A.; and Perez-Cruz, F., eds., *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, 1513–1521. PMLR.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. VizWiz Grand Challenge: Answering Visual Questions from Blind People. arXiv:1802.08218.

Hebbalaguppe, R.; Baranwal, M.; Anand, K.; and Arora, C. 2024. Calibration Transfer via Knowledge Distillation. In *Proceedings of the Asian Conference on Computer Vision*, 513–530.

Hebbalaguppe, R.; Kandar, T.; Nagpal, A.; and Arora, C. 2025. Prompting without Panic: Attribute-Aware, Zero-Shot, Test-Time Calibration. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 289–305. Springer.

Hebbalaguppe, R.; Prakash, J.; Madan, N.; and Arora, C. 2022. A stitch in time saves nine: A train-time regularizing loss for improved neural network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16081–16090.

Hu, Z.; Yang, P.; Li, B.; and Wang, Z. 2024. Multi-agents based on large language models for knowledge-based visual question answering. *arXiv preprint arXiv:2412.18351*.

Islam, M.; Seenivasan, L.; Ren, H.; and Glocker, B. 2021. Class-distribution-aware calibration for long-tailed visual recognition. *arXiv preprint arXiv:2109.05263*.

Jiang, B.; Zhuang, Z.; Shivakumar, S. S.; Roth, D.; and Taylor, C. J. 2024. Multi-agent vqa: Exploring multi-agent foundation models in zero-shot visual question answering. *arXiv preprint arXiv:2403.14783*.

Kull, M.; Perello Nieto, M.; Kängsepp, M.; Silva Filho, T.; Song, H.; and Flach, P. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32.

Kumar, A.; Sarawagi, S.; and Jain, U. 2018. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, 2805–2814. PMLR.

Lau, J. J.; Gayen, S.; Ben Abacha, A.; and Demner-Fushman, D. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1): 1–10.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, R.; Xu, C.; Guo, Z.; Fan, B.; Zhang, R.; Liu, W.; Zhao, Y.; Gong, W.; and Wang, E. 2022. AI-VQA: visual question answering based on agent interaction with interpretability. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5274–5282.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–3252. Dublin, Ireland: Association for Computational Linguistics.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lin, Z.; Zhang, D.; Tao, Q.; Shi, D.; Haffari, G.; Wu, Q.; He, M.; and Ge, Z. 2023. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143: 102611.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521.

Marcu, A.-M.; Chen, L.; Hünermann, J.; Karnsund, A.; Hanotte, B.; Chidananda, P.; Nair, S.; Badrinarayanan, V.; Kendall, A.; Shotton, J.; Arani, E.; and Sinavski, O. 2024. LingoQA: Visual Question Answering for Autonomous Driving. arXiv:2312.14115.

Mozaffari, M.; Sapkota, H.; and Yu, Q. 2025. GLEN: Generalized Focal Loss Ensemble of Low-Rank Networks for Calibrated Visual Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19563–19571.

Mukhoti, J.; Kulharia, V.; Sanyal, A.; Golodetz, S.; Torr, P.; and Dokania, P. 2020. Calibrating deep neural networks using focal loss. *Advances in neural information processing systems*, 33: 15288–15299.

Oksuz, K.; Kuzucu, S.; Joy, T.; and Dokania, P. K. 2023. Mocae: Mixture of calibrated experts significantly improves object detection. *arXiv preprint arXiv:2309.14976*.

Oriol, M.; Motger, Q.; Marco, J.; and Franch, X. 2025. Multi-Agent Debate Strategies to Enhance Requirements Engineering with Large Language Models. arXiv:2507.05981.

Pakdaman Naeini, M.; Cooper, G.; and Hauskrecht, M. 2015. Obtaining Well Calibrated Probabilities Using Bayesian Binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Patra, R.; Hebbalaguppe, R.; Dash, T.; Shroff, G.; and Vig, L. 2023. Calibrating Deep Neural Networks Using Explicit Regularisation and Dynamic Data Pruning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1541–1549.

Press, O.; Zhang, M.; Min, S.; Schmidt, L.; Smith, N. A.; and Lewis, M. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Qian, T.; Chen, J.; Zhuo, L.; Jiao, Y.; and Jiang, Y.-G. 2024. NuScenes-QA: A Multi-modal Visual Question Answering Benchmark for Autonomous Driving Scenario. arXiv:2305.14836.

Schulze, M.; Ebert, N.; Reichardt, L.; and Wasenmüller, O. 2025. Classifier Ensemble for Efficient Uncertainty Calibration of Deep Neural Networks for Image Classification. *arXiv preprint arXiv:2501.10089*.

Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Beißwenger, J.; Luo, P.; Geiger, A.; and Li, H. 2025. DriveLM: Driving with Graph Visual Question Answering. arXiv:2312.14150.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Tian, K.; Mitchell, E.; Zhou, A.; Sharma, A.; Rafailov, R.; Yao, H.; Finn, C.; and Manning, C. D. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Wang, Z.; Wan, W.; Lao, Q.; Chen, R.; Lang, M.; Wang, K.; and Lin, L. 2023a. Towards top-down reasoning: An explainable multi-agent approach for visual question answering. *arXiv preprint arXiv:2311.17331*.

Wang, Z.; Wan, W.; Lao, Q.; Chen, R.; Lang, M.; Wang, K.; and Lin, L. 2023b. Towards top-down reasoning: An explainable multi-agent approach for visual question answering. *arXiv preprint arXiv:2311.17331*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Whitehead, S.; Petryk, S.; Shakib, V.; Gonzalez, J.; Darrell, T.; Rohrbach, A.; and Rohrbach, M. 2022. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, 148–166. Springer.

Wieczorek, T. J.; Daun, N.; Khan, M. E.; and Rohrbach, M. 2025. Variational Visual Question Answering. *arXiv preprint arXiv:2505.09591*.

Yang, R.; Rajagopal, D.; Hayati, S. A.; Hu, B.; and Kang, D. 2024. Confidence calibration and rationalization for llms via multi-agent deliberation. *arXiv preprint arXiv:2404.09127*.

Yu, W.; Iter, D.; Wang, S.; Xu, Y.; Ju, M.; Sanyal, S.; Zhu, C.; Zeng, M.; and Jiang, M. 2022. Generate rather than retrieve: Large language models are strong context generators. *arXiv preprint arXiv:2209.10063*.

Zhang, X.; He, J.; Zhao, J.; Hu, Z.; Yang, X.; Li, J.; and Hong, R. 2024. Exploring and exploiting model uncertainty for robust visual question answering. *Multimedia Systems*, 30(6): 348.

Zhong, M.; Wang, G.; Chuang, Y.-N.; and Zou, N. 2025. Quantized Can Still Be Calibrated: A Unified Framework to Calibration in Quantized Large Language Models. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 30503–30517. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Zhou, Y.; Mei, J.; Yu, Y.; and Syeda-Mahmood, T. 2023. Medical visual question answering using joint self-supervised learning. arXiv:2302.13069.