CLVL 2025 Solution: Reliable VQA via Self-Reflection and Cross-Model Verification

Xixian Wu
Bilibili Inc.
Shanghai, China
wuxixian@bilibili.com

Yang Ou
Bilibili Inc.
Shanghai, China
ouyang10@bilibili.com

Pengchao Tian
Bilibili Inc.
Shanghai, China
tianpengchao@bilibili.com

Zian Yang
Bilibili Inc.
Shanghai, China
yangzian@bilibili.com

Jielei Zhang
Bilibili Inc.
Shanghai, China
zhangjielei@bilibili.com

Abstract

Visual Question Answering (VQA) represents a key research area at the intersection of visual perception and linguistic reasoning, which has been made significant progress in recent years. However, due to the opaque and uncontrollable reasoning processes in vision-language models (VLMs), the capacity of these models to evaluate the correctness of their responses remains largely underexplored. To address this, we propose a novel approach that integrates Self-Reflection and Cross-Model Verification to comprehensively assess the uncertainty of model responses. We first introduce two selector models that harnesses the latent representations of the VLM alongside embeddings of the corresponding VQA pairs to estimate the reliability of each response. Furthermore, to mitigate the inherent hallucinations of the VLM, we incorporate reference models to verify the correctness of VLM's answers. In the test phase of the Reliable VQA Challenge at ICCV-CLVL 2025, our method achieves a Φ_{100} *score of 39.64 and a 100-AUC of 97.22.*

1. Introduction

Visual Question Answering (VQA) is a fundamental task in multi-modal understanding, where a model is required to answer questions based on the content of an image. Despite achieving remarkable performance on VQA benchmarks, vision-language models (VLMs) remain difficult to deploy in safety-sensitive downstream applications due to their opaque and uncontrollable reasoning processes.

One way to address this issue is to formulate the problem as selective prediction. In this paradigm, a model is not only required to produce an output but also to assesse the reliability of its prediction and abstains when uncertain. In the context of the Reliable VQA Challenge, a sub-challenge of the ICCV-CLVL 2025 workshop, participants are required to not only generate answers to VQA tasks but also provide a confidence score for each response. The evaluation additionally requires specifying a global abstention threshold, under which responses are regarded as uncertain and consequently treated as abstentions.

Recently, several studies have been proposed to solve such problem, [4] explores the effectiveness of several selection functions including MaxProb, Calibration and Multimodal selection function. [2] propose Learning from Your Peers (LYP) approach for training multimodal selection functions for making abstention decisions. Other efforts [1, 3] have attempted to enable VLMs to assess the reliability of their own answers. However, due to the intrinsic hallucination problem of VLMs, the model may still exhibit overconfidence in its erroneous predictions.

In this study, we propose a selective prediction framework that integrates model self-reflection and cross-model verification, enabling the system to assess its own reliability and leverage consensus across models to mitigate overconfident or hallucinated responses. To summarize, our contribution can be expressed as follows:

- We introduce a cross-model verification mechanism on top of model self-reflection, effectively mitigating the intrinsic hallucination problem of vision-language models.
- We leverage an ensemble of multiple models to enhance the estimation of prediction reliability, resulting in substantial improvements.
- Extensive evaluation demonstrates that our approach achieves state-of-the-art performance on the Reliable VQA Challenge leaderboard.

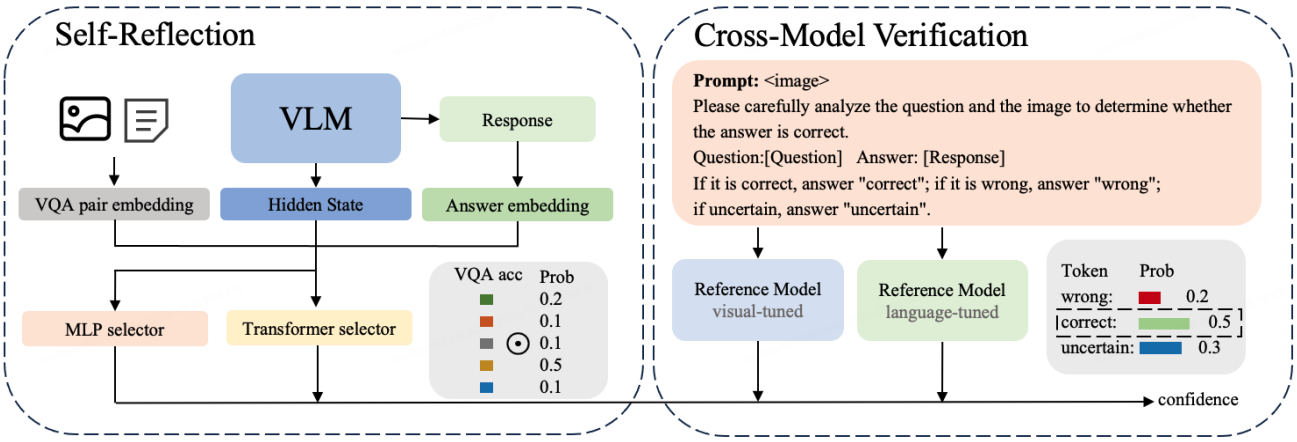

Figure 1. Overall framework of the proposed method.

2. Method

The overall framework of our approach is illustrated in Figure 1. We first assess the reliability of the VLM's outputs using its internal representations. Then, we validate the answers with external reference models and generate robust predictions through model ensembling.

2.1. Self-Reflection

To assess the model's uncertainty in its own predictions, we extract the last-layer hidden states from the VLM during inference for each VQA image-text pair. These representations inherently encode the rich multimodal interactions between visual and linguistic inputs [5]. In addition, we extract textual embeddings of both the question and the model-generated answer using CLIP, providing complementary semantic cues to the multimodal representations. Considering that the VQA accuracy is a discrete value, we formulate the problem as a classification task, which tends to be more stable and easier to optimize than a regression formulation.

Formally, we consider a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where y_i represents the discretized VQA accuracy serving as a categorical label. Each input x_i consists of the VLM's hidden state, the embedding of the question, and that of the answer, i.e.,$x_i = [h_i; q_i; a_i]$, where h_i denotes the hidden state representation, q_i the textual feature of the question, and a_i the textual feature of the answer.

The learning objective is formulated as a standard multi-class classification problem. Given the categorical label y_i, the model is trained to learn the mapping from x_i to y_i by minimizing the cross-entropy loss. Formally, the loss function is defined as:

$$\mathcal{L}_{\text{CE}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=0}^{C-1}\mathbb{I}(y_i = c)\log p_\theta(y_i = c \mid x_i), \quad (1)$$

with C denoting the total number of classes and $\mathbb{I}(\cdot)$ representing the indicator function. This objective encourages the predicted probability distribution $p_\theta(y_i \mid x_i)$ to be closely aligned with the ground-truth label distribution.

In practice, due to the relatively low proportion of negative samples, we adopt the focal loss to address this imbalance issue and improve the model's learning on hard examples. The focal loss is formulated as follows:

$$\mathcal{L}_{\text{FL}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=0}^{C-1}\alpha_c\,(1 - p_\theta(y_i = c \mid x_i))^\gamma \times \mathbb{I}(y_i = c)\,\log p_\theta(y_i = c \mid x_i), \quad (2)$$

where $\alpha_c \in [0, 1]$ is a weighting factor for class c to address class imbalance, and $\gamma > 0$ is the focusing parameter that emphasizes hard-to-classify examples.

At inference time, we estimate the confidence score $\hat{s}_i$ by taking the expectation of the discretized accuracy values with respect to the predicted categorical distribution:

$$\hat{s}_i = \sum_{c=0}^{C-1} a_c\, p_\theta(y_i = c \mid x_i), \quad (3)$$

where a_c denotes the VQA accuracy associated with class c. This formulation allows the model to output a continuous confidence estimate based on the learned discrete probability distribution.

Based on the above, we independently train two models, one using an MLP architecture and the other using a Transformer architecture. For each sample, the resulting confidence scores are denoted as $\hat{s}_i^{\text{MLP}}$ and $\hat{s}_i^{\text{T}}$, respectively.

2.2. Cross-Model Verification

The selector heavily relies on the quality of the VLM hidden states. To mitigate this limitation, we introduce an ex-

Table 1. Comparison with other methods.

Team	Accuracy	Φ_{100}	100-AUC	Cov@0.5%	Cov@1% (%)	Cov@2%	Cov@5%
jokur (Test-standard)	84.05	32.06	96.85	40.67	49.30	59.15	75.30
Baseline (Test-standard)	81.35	23.00	96.09	26.98	39.75	52.88	71.78
Ours (Test-dev)	83.99	37.52	97.15	44.56	53.03	63.49	77.97
Ours (Test-standard)	**84.11**	**39.64**	**97.22**	**45.67**	**54.12**	**63.71**	**78.55**

Table 2. performance on the VQAV2 validation set.

Model	Accuracy (%)
QwenVL2-7B-Inst	82.97
QwenVL2-7B-Inst CoT	65.91
InternVL3_5-8B	74.08
InternVL3_5-8B-Inst	78.02
InternVL3_5-GPT-OSS-20B	80.06
InternVL3_8B-Inst	79.79
QwenVL2.5-7B-Inst	76.23
QwenVL2-7B-Inst (TTA)	**83.14**

Table 3. Ensemble of Models.

Team	Φ_{100}	100-AUC	Cov@0.5%
MLP	30.91	96.67	37.79
MLP+T	31.52	96.7	37.94
MLP+T+V	36.31	97.03	43.13
MLP+T+V+L	**39.64**	**97.22**	**45.67**

ternal reference model to verify the answers generated by the VLM.

Specifically, given the original VQA pair and the answer generated by the VLM, we use Qwen2.5-VL-7B-Inst to assess whether the answer is *correct*, *wrong*, or *uncertain*. We derive a confidence score from the probability that Qwen2.5-VL generates the word `"correct"`:

$$\hat{s}_i^{\text{VLM}} = p_{\text{Qwen}}(\text{"correct"} \mid x_i, \text{VLM answer}), \quad (4)$$

where $p_{\text{Qwen}}(\cdot)$ is the predicted probability assigned by the external reference model to its first output token.

To improve the performance of Qwen2.5-VL, we fine-tuned the model by separately updating its visual and language modules. For training, we categorized the VLM-generated answers based on their VQA accuracy: answers with a VQA accuracy of 0 were labeled as *wrong*, those with accuracy greater than 0 but less than or equal to 0.66 were labeled as *uncertain*, and answers with accuracy greater than 0.66 were labeled as *correct*. We then fine-tuned Qwen2.5-VL using supervised fine-tuning (SFT) on this labeled dataset to enhance its ability to accurately verify VLM answers.

After fine-tuning, we perform inference with the visual-tuned and language-tuned models to obtain confidence scores for each sample. The resulting scores are denoted as $\hat{s}_i^{\text{VLM-V}}$ and $\hat{s}_i^{\text{VLM-L}}$, respectively, representing the model's estimated probability that a given VLM-generated answer is correct.

To this end, we progressively aggregate the outputs from all models to obtain the final confidence score, which is expressed as follows:

$$\hat{s}_i^{\text{MLP+T}} = 0.5\hat{s}_i^{\text{MLP}} + 0.5\hat{s}_i^{\text{T}}, \quad (5)$$

$$\hat{s}_i^{\text{MLP+T+V}} = 0.5\hat{s}_i^{\text{MLP+T}} + 0.5\hat{s}_i^{\text{VLM-V}}, \quad (6)$$

$$\hat{s}_i^{\text{MLP+T+V+L}} = 0.5\hat{s}_i^{\text{MLP+T+V}} + 0.5\hat{s}_i^{\text{VLM-L}}. \quad (7)$$

3. Experiments

3.1. Implement Details

We employ QwenVL2-7B-Inst for inference due to its strong performance on the VQAv2 dataset. To further enhance performance, we apply test-time augmentation (TTA) by increasing the temperature and aggregating multiple inference results, which leads to a slight improvement in overall accuracy.

For the MLP selector, the question and answer features are each processed through one fully connected layer, while the hidden state features pass through two fully connected layers. The resulting features are concatenated and fed into three fully connected layers for classification. In contrast, for the Transformer selector, a two-layer Transformer encoder is used to model interactions among the three features, and classification is performed based on the hidden state token. The checkpoint achieving the best performance on the validation set is used for testing.

For the external reference model Qwen2.5-VL, we fine-tune its visual module by unfreezing the multi-modal projector, while the language module is fine-tuned using the LoRA strategy. All fine-tuning processes are conducted for one epoch, and the final checkpoint is used for testing.

The abstention threshold is determined by sweeping across different threshold values and selecting the one that yields the best performance on the validation set.

3.2. Main Results

The final results are presented in the table 1, including only the public leaderboard entries and the performance of our method. While the accuracy is comparable to existing approaches, our method consistently outperforms them across

all other metrics, achieving a Φ_{100} of 39.64, a 100-AUC of 97.22, and a Cov@0.5% of 45.67, among others. Especially, in terms of Φ_{100}, our method outperforms the second-best approach by 7.58.

3.3. Ablation Study

Performance of VLMs on the VQAV2 dataset: From Table 2, we observe that Qwen2 achieves the best performance on the VQAv2 dataset, even surpassing its subsequent versions. Moreover, applying TTA can slightly enhance the performance of the VLM.

Ensemble of Models: We report the performance of MLP+T, MLP+T+V, and MLP+T+V+L in Table 3. The results indicate that each model contributes positively to the prediction, demonstrating the effectiveness of our approach.

4. Conclusion

In this work, we present a novel selective prediction framework for vision-language models that combines model self-reflection with cross-model verification. By enabling models to internally assess the reliability of their own predictions and to leverage consensus across multiple models, our approach effectively mitigates overconfident and hallucinated responses, a major challenge in deploying VLMs in safety-critical applications. Extensive experiments on the Reliable VQA Challenge demonstrate that our method not only improves the trustworthiness of predictions but also establishes a new state-of-the-art performance on the leaderboard. In the future, we plan to extend this framework to broader multi-modal reasoning tasks and explore more efficient mechanisms for model collaboration and uncertainty estimation.

References

[1] Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan O Arik, Tomas Pfister, and Somesh Jha. Adaptation with self-evaluation to improve selective prediction in llms. *arXiv preprint arXiv:2310.11689*, 2023. 1

[2] Corentin Dancette, Spencer Whitehead, Rishabh Maheshwary, Ramakrishna Vedantam, Stefan Scherer, Xinlei Chen, Matthieu Cord, and Marcus Rohrbach. Improving selective visual question answering by learning from your peers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24049–24059, 2023. 1

[3] Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Raghavi Chandu. Selective" selective prediction": Reducing unnecessary abstention in vision-language reasoning. *arXiv preprint arXiv:2402.15610*, 2024. 1

[4] Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, pages 148–166. Springer, 2022. 1

[5] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification?, 2024. 2