Cross-Modal Grounding for Calibrated Confidence in Vision Language Models

Trilok Padhi[1*†] **Ramneet Kaur**[2*] **Adam D. Cobb**[2] **Manoj Acharya**[2]
Colin Samplawski[2] **Brian Matejek**[2] **Alexander M. Berenbeim**[3]
Nathaniel D. Bastian[3] **Susmit Jha**[2] **Ugur Kursuncu**[1] **Anirban Roy**[2]
[1]Georgia State University, Atlanta, USA
[2]Computer Science Lab, SRI, Menlo Park, USA
[3]Army Cyber Institute, United States Military Academy, West Point, NY, USA
tpadhi1@student.gsu.edu, ramneet.kaur@sri.com

Abstract

We introduce a novel approach for calibrating uncertainty quantification (UQ) tailored for visual language models (VLMs). Existing state-of-the-art UQ methods rely on consistency among multiple responses, but often report higher confidence when the model is consistently incorrect, leading to poorly calibrated confidence. To address this, we leverage cross-modal consistency by grounding back the textual responses to visual inputs. The confidence from the grounding model (GM) is used to calibrate the overall confidence. Given that using a GM adds its own uncertainty, we apply temperature scaling to calibrate the GM's confidence in the accuracy of the generated responses. We comprehensively evaluate the proposed approach across diverse range of settings, including datasets with varying visuo-linguistic complexity, Q&A and caption generation tasks, general visual and specialized medical contexts, VLMs differing in scale, openness, and different categories of GMs. Experimental results show that our approach consistently improves confidence calibration, with expected calibration error (ECE) improvements ranging from 32.5% to 96.7% over both self and non-self consistency-based baselines. Code is available here https://github.com/trilokpadhi/Multimodal-Uncertainty-Quantification.

1 Introduction

Recent advancements in large language models (LLMs) have demonstrated their impressive performance across many domains, ranging from natural language processing (Devlin et al., 2019) and machine translation (Chitale et al., 2024) to creative writing (Gómez-Rodríguez and Williams, 2023) and code generation (Jiang et al., 2024). Despite their capabilities, these models are not infallible and are known to produce incorrect or misleading information, often referred to as hallucinations (Huang et al., 2025). Uncertainty Quantification (UQ) of LLMs has been proposed as a practical solution to assess trust in these models, particularly for their deployment in safety-critical areas such as healthcare (Shorinwa et al., 2024). UQ techniques aim to provide a quantitative measure of trust that a user can place in an LLM's response to the input query.

State-of-the-art (SOTA) approaches for quantifying the uncertainty of LLMs are motivated by self-consistency theory (Wang et al., 2022). This involves prompting the model multiple times for the same input under diverse settings, such as with a high temperature, and checking for similarity in the generated responses for assessing model's uncertainty on the input (Lin et al., 2023; Kuhn et al., 2023; Kadavath et al., 2022; Kaur et al., 2024). The underlying idea is that if the model generates semantically similar responses for the same input under diverse settings, then it is certain (or confident) about their input.

Consistency, however, does not imply accuracy. As shown in Fig. 1 (left), we observe that models can generate consistently incorrect responses. This observation is made on 25 inputs out of 75 manually verified test cases on Slake, a medical dataset (Liu et al., 2021). The goal of UQ is to provide a measure of confidence in the real-world performance of LLMs for their trustworthy deployment. The confidence in an LLM as reported by UQ approaches should, therefore, be aligned with the accuracy of the model. This alignment can be checked by plotting the expected accuracy of the model as a function of the reported confidence. These plots are known as *reliability diagrams* and have been used to report confidence-accuracy calibration (or alignment) of deep learn-

[*]Equal contribution.
[†]Trilok Padhi is a graduate student in the Department of Computer Science at Georgia State University, Atlanta, USA. This work was done when he was a summer intern at SRI.

Proceedings of the 1st Workshop on Grounding Language Models: Learning Faithfully and Efficiently (GroundLM 2026), pages 480–504
October 29, 2026 ©2026 Association for Computational Linguistics

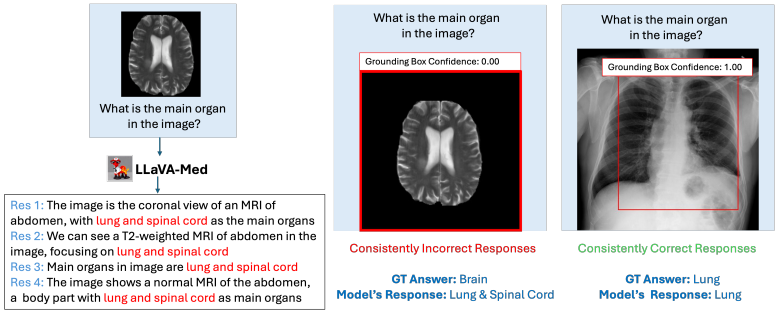


Figure 1: Consistently incorrect responses (Res 1..Res 4) by LLaVA-Med-v1.5-Mistral-7B (Li et al., 2023a) VLM on MRI image of brain from the Slake Medical dataset (left). BiomedParse (Zhao et al., 2024), a grounding model for medical images, is not able to locate 'lung and spinal cord' on the MRI image of brain, and therefore labels the entire image as lung with 0 confidence. It is, however, able to generate a bounding box for lung on the chest X-ray with 100% confidence (right).

ing models (Guo et al., 2017).

Fig. 4a and Fig. 4b in Appendix show reliability diagrams for two self-consistency-based UQ approaches for predicting confidence of Llama-2-13B (Touvron et al., 2023) on TriviaQA (Joshi et al., 2017) and COQA (Reddy et al., 2019) datasets respectively; a common test setting considered by these approaches (Lin et al., 2023; Kuhn et al., 2023; Kaur et al., 2024). High expected calibration error (ECE) shows poor calibration of the model's accuracy with its confidence predicted by these approaches We observe similar trends of high ECE by UQ approaches when applied to multi-modal input settings of text-image pairs for these models. These observations on visual language models (VLMs) are reported in the experimental section.

In this paper, we propose a cross-modal grounding approach for calibrating confidence in VLMs. The goal is to calibrate the confidence estimates produced by existing uncertainty quantification (UQ) methods by leveraging cross-modal consistency, specifically by grounding the VLM's textual responses back into the visual modality of the input image. For example, as shown in Fig. 1 (right), we ground the answers with bounding boxes on the image. A correct answer has a higher chance of being grounded as the evidence is likely to be present in the image. Thus, the ability (or inability) to ground the generated response to the input image provides evidence about the correctness (or incorrectness) of the response.

One consideration of relying on a grounding model (GM) to report the uncertainty of another model is that it introduces the GM's uncertainty into the pipeline. We apply temperature scaling to calibrate the GM's confidence in the accuracy of the VLM. Temperature scaling is a simple yet effective post-processing parametric calibration technique that has been widely used to align the confidence with the true likelihood of correctness (Jaynes, 1957; Hinton et al., 2015; Guo et al., 2017). Fig. 2 shows our proposed approach for calibrating confidence in predictions by VLMs. The contributions of this paper can be summarised as:

1. Miscalibration of UQ Approaches: We observe and report that existing UQ approaches for LLMs as well as VLMs exhibit poor calibration between the reported confidence and the model's accuracy.

2. Multi-modal Consistency: We propose a novel method to calibrate confidence in vision-language models (VLMs) by leveraging cross-modal consistency between the generated textual responses and the input image.

3. Significant Calibration Improvements: We show that our approach is generalizable by demonstrating significant calibration improvements in ECE (upto 96.7%) with experiments across *VLMs*: varying in size and type, *domains*: general, medical, reasoning, *datasets*: VQA, Slake, GQA, Whoops, *tasks*: Q&A, and caption generation, *grounding models*: semantic-based, segmentation-based, and foundation models, *UQ approaches*: SOTA self-consistency-based as well as non-self-consistency-based verbalized prompting approaches.

2 Background

2.1 Uncertainty Quantification

Uncertainty quantification (UQ) techniques aim to measure the uncertainty in the predictions of LLMs to assess their reliability. A prominent strategy for UQ is built on self-consistency theory (Wang et al., 2022), which focuses on generating multiple responses from the model for the same input and evaluating the consistency among these outputs.

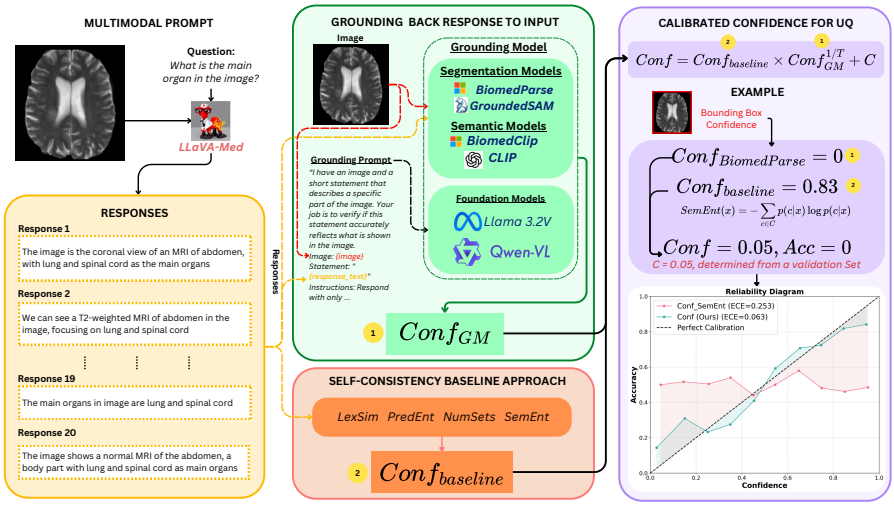


Figure 2: $Conf_{baseline}$: confidence from a self-consistency UQ baseline such as LexSim (Fomicheva et al., 2020), PredEnt (Malinin and Gales), NumSets (Kuhn et al., 2023), SemEnt (Kuhn et al., 2023), etc. on a VLM such as LLaVA-Med. $Conf_{GM}^{1/T}$: temperature-scaled calibrated confidence of grounding model on the accuracy of the generated responses. A grounding model can be as simple as the CLIP-based model, that provides its confidence in terms of a similarity score between embeddings of the generated response and the input image (e.g. BiomedCLIP), a detection model for response on the input image reporting its confidence on the detected bounding box (e.g. BiomedParse), or a foundation model that provides its verdict – a confidence score in $[0, 1]$ – on the relevance of the generated response to the input image (e.g. LLaMA 3.2V, Qwen VL, etc.). $Conf$ (2) is the proposed calibrated confidence score for UQ of VLMs resulting in substantially lower ECE than the baseline.

This strategy quantifies uncertainty by identifying discrepancies in the generated responses. Different approaches for UQ of LLMs differ in how they measure these discrepancies and the metrics they use to quantify the resulting uncertainty.

Predictive entropy over the probability distribution of responses is a popular uncertainty metric, and serves as a baseline in UQ for LLMs (Kadavath et al., 2022; Kuhn et al., 2023; Kaur et al., 2024; Lin et al., 2023). In the context of natural language processing, for an input query x the probability of a response r is calculated as the product of the conditional log probabilities for each token in the response: $p(\mathbf{r}|x) = \prod_i p(r_i|r_{<i}, x)$. *Lexical similarity* (Fomicheva et al., 2020) is another proposed approach that assigns a similarity score between each pair of responses (r_i, r_j) via RougeL (Lin, 2004)[1], and uses the average of this score over each pair as the confidence (1-UQ) metric in the LLM: $\frac{1}{P}\sum_{i=1}^{|R|}\sum_{j=1}^{|R|} \text{RougeL}(r_i, r_j)$. Here R is the set of responses, and $P = |R| \times (|R|-1)/2$. Kuhn et al. (2023) group the generated responses into semantically equivalent response clusters C, compute the probability of each cluster as the average probability of each response in the cluster, and use *semantic entropy* (SE) over these clusters as the uncertainty metric: $SE(x) = -\sum_{c \in C} p(c|x) \log p(c|x)$, Semantic equivalence between two responses for clustering them together is checked via bi-directional entailment between the two via Deberta (Natural Language Inference) model (He et al., 2020). Kaur et al. (2024) further enhances this approach by a new dynamic semantic clustering algorithm for deciding on the membership of a response. Another UQ metric as a baseline in (Kuhn et al., 2023; Kaur et al., 2024) is *NumSets*, that is the number of semantic clusters formed by the responses.

2.2 Confidence Calibration

Confidence in a model's predictions is known to be calibrated if it represents the true probability of those predictions to be correct (Guo et al., 2017). This means that for a set of predictions where the model's confidence estimate is c, the proportion of correct predictions by the model should be c.

Temperature Scaling (Guo et al., 2017) is a simple yet effective post-processing parametric method for calibrating the softmax confidence of classification models. It learns a single parameter (T) to scale the model's logits by a factor of $1/T$, with greater values of T softening the softmax distribution indicating high uncertainty, and lower values of T sharpening the distribution indicating low uncertainty.

Reliability diagrams (DeGroot and Fienberg, 1983), such as the ones shown in Fig. 4b, provide

[1]RougeL measures the similarity between two sentences based on the longest common subsequence between the two.

us with a way to visualize confidence calibration empirically. These diagrams plot the expected accuracy of the model as a function of confidence. The range of confidence in $[0, 1]$ is divided into equal-sized smaller bins, and accuracy for each bin is calculated as an average accuracy on the samples falling in that bin. A perfect calibration would result in average accuracy for each bin equal to the average confidence of the bin. Any deviation from this identity function indicates a miscalibration. *Expected Calibration Error* (ECE) (Naeini et al., 2015) provides us with a quantitative measure for this miscalibration. ECE is calculated as the weighted average over each bin's difference in accuracy and confidence (Guo et al., 2017):

$$\sum_{m=1}^{M} \frac{N_m}{n} |Acc_m - Conf_m|, \quad (1)$$

where M is the number of confidence bins, N_m is the number of samples in bin m, n is the total number of samples, and Acc_m, $Conf_m$ are the average accuracy and confidence of the m^{th} bin.

2.3 Grounding

Grounding refers to the process of linking symbolic representations of knowledge (e.g. language) to the real world's sensory data (e.g., images, sounds). For example, visual grounding techniques (Yu et al., 2018; Acharya et al., 2019) have been proposed for mapping textual entities to bounding boxes on images. Grounding models (GM) such as object detection or localization models have been used to report interpretability of LLMs (Giulivi and Boracchi, 2024; Sahu et al., 2024). We propose the use of grounding models for calibrating the UQ of VLMs. Different types of GM can be utilized for linking back the generated response by an LLM to the input space. We consider three of those: *segmentation-based GM* that generates a segmentation mask relevant to the generated response on the input image, *semantic-based GM* that assigns a similarity score to the generated response and the input image, and a *foundation model* that can be instructed to verify the accuracy of the response as the description of the input image.

3 Related Work

There has been recent work on utilizing information from different modalities of multi-modal inputs to VLMs for their interpretability. Giulivi and Boracchi (2024) make use of a localization model (OWL-ViT) for projecting bounding box on the identified objects by the VLM. They do so by a joint training in the embedding space of the LLM and OWL-ViT. This approach is applicable to open-source models for joint training in the embedding space of vision modality. Sahu et al. (2024) run object detection model on an input image and make use of the detected objects to check the LLM's response for hallucinations. This is done by generating a claim from the query-response pair and then verifying the decomposed sub-claims against the detected objects.

There has been a recent focus on measuring uncertainty in VLMs via self-consistency theory. Zhang et al. (2024) apply perturbations in both text and image input modalities, and use entropy in the distribution of the generated responses for reporting uncertainty of the LLM on the input query. This can be used as the self-consistency UQ baseline in the proposed framework for calibrating the reported confidence (by the baseline) in VLMs. Li et al. (2024) propose to train a Graph Neural Network (GNN) on clusters of semantically equivalent responses to predict the probability of each response being correct, with the optimization goal of calibrating the predicted probabilities. This is, however, a supervised approach that requires a labeled training set for the GNN. Our approach does not require training any new model but involves calibrating the confidence of grounding models (GM) via temperature scaling on a small validation set to take into account the uncertainty of GM used in the pipeline. The use of a validation set is a common setting in uncertainty quantification literature such as conformal prediction (Balasubramanian et al., 2014), where the validation set is used to determine a threshold on the membership score (known as non-conformity score) of a generated response in the output set. The output is a set instead of a single prediction from the LLM to take into account uncertainty in the LLM with coverage guarantees on the set (Wang et al., 2025; Ye et al., 2024; Kaur et al., 2024; Quach et al., 2023).

Prompting LLMs to generate a confidence score along with its prediction has also been used to quantify the model's uncertainty (Tian et al., 2023; Xiong et al., 2023). High ECE in the self-confidence by VLMs has reported in the literature (Kostumov et al., 2024; Groot and Valdenegro-Toro, 2024). The authors report overconfidence by VLMs but do not propose any confidence technique for those. To the best of our knowledge, this is the

first work on calibrating confidence from existing UQ approaches when applied to VLMs.

4 Calibrated Uncertainty Quantification

The core idea behind the proposed approach is to leverage cross-modal consistency for evaluating the accuracy of the generated responses. This inference time accuracy evaluation can be utilized for calibrating the UQ of VLMs. The ability to ground back a response into the provided context generates evidence about the response's accuracy. Fig. 2 shows examples of grounding models (GM) that quantitatively provide this evidence by estimating a confidence score in the accuracy of the VLM's response. We use this GM-generated confidence (a score in $[0, 1]$) to calibrate the confidence reported by the existing UQ approaches. Specifically, confidence in a VLM by a UQ baseline, $Conf_{baseline} \in [0, 1]$, is calibrated by multiplying it with the GM's confidence in the accuracy of the model on the input query:

$$Conf = Conf_{baseline} \times Conf_{GM}^{(1/T)} + C, \quad (2)$$

where $Conf_{GM}$ is the average confidence of the grounding model over multiple responses generated for the self-consistency UQ baseline.[2] $T(> 0)$ is the hyper-parameter for temperature scaling or calibration of $Conf_{GM}$. Higher values of T sharpen (or increase) the confidence of the grounding model. Reducing the value of T reduces the confidence of the grounding model, and therefore the overall confidence. This is necessary to make sure the grounding model is calibrated correctly to the task. A constant $C(> 0)$ is also added to offset the reduction caused by the product of confidences in the range of $[0, 1]$. Both hyperparameters T and C are determined from a validation set.

Specifically, Conf_{GM} is computed as follows depending on the GM type: **Semantic-based (CLIP/BiomedCLIP):** cosine similarity between the text response and image embeddings, normalized to $[0, 1]$; **Segmentation-based (Grounded-SAM):** bounding-box detection confidence, natively in $[0, 1]$; **Foundation-based (LLaMA3.2V, QwenVL):** we assign a score of 1 for `Yes` responses and 0 for `No` or `Maybe` responses. In all cases, $Conf$ is clipped to 1 when C pushes the sum above 1.0. For foundation GMs, the grounding prompt asks the model to verify whether the textual response accurately reflects the input image, responding with only `Yes`, `No`, or `Not sure`. We map to 0 to both `No` and `Not sure`, reflecting a deliberate conservative design choice: in safety-critical settings, the inability to confirm a fact is functionally equivalent to a failure to provide grounding evidence (see Appendix for full prompt details).

Theoretical Motivation. Equation (2) can be viewed as a structured correction for the primary failure mode of self-consistency-based uncertainty quantification: the assumption that response consistency necessarily implies correctness. We consider three cases.

Case 1 — Model Inconsistency. When the base model is inconsistent, $\text{Conf}_{\text{baseline}}$ is already low, which naturally yields a low final confidence score.

Case 2 — Consistent but Incorrect. This is the critical setting in which existing self-consistency baselines often fail. The grounding term $\text{Conf}_{\text{GM}}^{1/T}$ introduces evidence from the visual modality. If the grounding model cannot identify visual support for the textual response, Conf_{GM} approaches zero, thereby correcting an overconfident baseline towards a calibrated low-confidence prediction.

Case 3 — Consistent and Correct. When the response is both consistent and visually supported, both terms remain high, preserving justified high confidence. The additive offset C is motivated by the fact that multiplying two values in $[0, 1]$ inherently compresses the resulting score; for example, $0.8 \times 0.8 = 0.64$. Without C, the formulation would systematically underestimate confidence for correct predictions. The offset shifts the confidence distribution closer to the $x = y$ identity line in the reliability diagram, thereby reducing ECE.

The formulation can also be interpreted as the ***Product-of-Experts*** with two complementary experts: *semantic consistency* ($\text{Conf}_{\text{baseline}}$), and *visuo-linguistic grounding* (Conf_{GM}). The power transformation $1/T$ acts as a parametric weight that calibrates the influence of grounding evidence. Their product implements a conjunctive AND constraint: high final confidence requires the response to be both internally consistent and externally grounded. The additive offset C further prevents confidence from collapsing to zero due to limitations of the GM, while still preserving the VLM's own consistency signal.

[2] In case of non-consistency-based baselines, $Conf_{GM}$ is the grounding-model's confidence in a single response by the VLM.

5 Experiments

5.1 Datasets

We evaluate our method on two open-ended visual question answering (QA) benchmarks from different domains: the medical-domain **Slake** (Liu et al., 2021) dataset and the general-domain **VQA v2.0** (Goyal et al., 2017) dataset. Slake is a bilingual (English and Chinese) dataset containing 14K QA pairs over 642 CT, MRI, and X-ray images, with both vision-only and knowledge-based questions annotated by physicians. We use only the English QA pairs and create 80/20 test/validation splits from the original test set. VQAv2 evaluates general visuo-linguistic understanding while reducing reliance on language priors through balanced annotations; we similarly use an 80/20 test/validation split from its test set. Additional examples from both datasets are provided in the Appendix.

We also perform ablation studies on **GQA** (Hudson and Manning, 2019) and **Whoops** (Bitton-Guetta et al., 2023). While VQA primarily evaluates general visual understanding (e.g., object recognition, counting, color identification, and factual reasoning), GQA requires compositional, multi-step reasoning over scene graphs. Whoops further challenges vision–language models with adversarial questions designed to avoid shortcuts based on dataset biases or scene-graph reasoning. Example image–question pairs from GQA and Whoops are included in the Appendix.

5.2 Vision Language Models

We consider LLaVA-v1.5-7B (Liu et al., 2024a) and LLaVA-Med-v1.5-Mistral-7B (Li et al., 2023b) as the VLMs for quantifying their uncertainty on VQA and Slake, respectively. **L**arge **L**anguage **a**nd **V**ision **A**ssistant (LLaVA) is a vision and language model that connects a vision encoder (CLIP (Radford et al., 2021)) with a language model (LLaMA 2 (Touvron et al., 2023)) to handle image-text queries. LLaVA-Med-v1.5-Mistral-7B is the fine-tuned variant of LLaVA on the medical domain with Mistral-7B (Jiang et al., 2023) as the language model. The fine-tuning is done by curriculum learning on biomedical image-caption pairs from the PubMed Central (Zhang et al., 2023a) dataset. We also conduct ablation studies on different VLMs, varying in openness and scale. Details and results of these studies on those models are in Section J.

5.3 Grounding Models

We consider different categories of the grounding models (GM). For VQA, we consider the following GM: **1. Segmentation-based:** GroundedSAM (Ren et al., 2024) is trained to perform segmentation on the bounding box detected by Grounding DINO (Liu et al., 2024b) on an image for an input text. **2. Semantic-based:** The **C**ontrastive **L**anguage–**I**mage **P**re-training CLIP (Radford et al., 2021) model assigns a similarity score between an image-text pair in the shared embedding space. **3. Foundation-based:** LLaMA-3.2-11B-Vision-Instruct (Grattafiori et al., 2024) (LLaMA3.2V), and Qwen2-VL-7B-Instruct (Wang et al., 2024) (QwenVL). For Slake, we report results with semantic and foundation GM: **1. Semantic-based GM:** BiomedClip (Zhang et al., 2023b), an advanced version of CLIP model fine-tuned on the medical domain. **2. Foundation GM:** Biomed-Qwen2-VL-2B-Instruct (Cheng et al., 2024) (Biomed-QwenVL), the fine-tuned version of QwenVL on the medical domain, and LLaMA-3.2-11B-Vision-Instruct[3]. We add prompt details for foundation GM in the Appendix.

5.4 Baselines

We consider all four self-consistency-based UQ baselines covered in the background section: Predictive Entropy (PredEnt), Lexical Similarity (LexSim), Semantic Entropy (SemEnt), and NumSets. For each prompt, we elicit 20 responses from VLM in order to capture diverse settings for the LLM with temperature $= 0.5$ for randomness, and top_p $= 1$ for nucleus sampling. For a fair comparison, we also consider calibrating all baselines directly with temperature scaling ($Conf_{baseline}^{(1/T)}$) where T is learned using the validation set.

We also conduct ablations on LLaVA for VQA on all four non-consistency-based verbalized prompting UQ approaches recently from Xuan et al. (2025). In these methods, models are explicitly instructed to output confidence scores for their generated responses, either through one-stage or two-stage prompting. One-stage prompting produces both the answer and its confidence within a single generation. This includes *Chain-of-Thought (CoT)* (Wei et al., 2022), where the model generates step-by-step reasoning followed by an answer and

[3]To the best of our knowledge, there is no model LLaMA3.2V family of foundation models specific to the medical domain

Baseline	Without Grounding (↓)		With Grounding (↓)			
	$Conf_{\text{baseline}}$	$Conf^{(1/T)}_{\text{baseline}}$	GroundedSAM	CLIP	LLaMA3.2V	QwenVL
LexSim	0.169	0.081	0.052 (−36.0%)	**0.045** (−44.6%)	**0.045** (−44.6%)	0.061 (−25.0%)
NumSets	0.410	0.246	0.143 (−41.9%)	0.143 (−41.8%)	0.127 (−48.4%)	**0.125** (−49.2%)
PredEnt	0.445	0.190	0.115 (−39.6%)	0.117 (−38.3%)	**0.096** (−49.4%)	0.119 (−37.5%)
SemEnt	0.108	0.108	0.036 (−66.9%)	0.038 (−64.6%)	**0.029** (−73.1%)	0.073 (−32.5%)

Table 1: Comparison of ECE over accuracy of LLaVA for VQA with the confidence reported by baseline ($Conf_{baseline}$), calibrated baseline ($Conf^{(1/T)}_{baseline}$), and the proposed calibration with grounding 2. Percentage improvement in ECE via grounding from the calibrated baseline is also reported (in green) for each grounding model.

Baseline	Without Grounding (↓)		With Grounding (↓)		
	$Conf_{\text{baseline}}$	$Conf^{(1/T)}_{\text{baseline}}$	BiomedClip	LLaMA3.2V	Biomed-QwenVL
LexSim	0.031	0.031	**0.004** (−87.1%)	0.013 (−58.1%)	0.038 (+22.6%)
NumSets	0.426	0.201	**0.007** (−96.5%)	0.021 (−89.6%)	0.132 (−34.3%)
PredEnt	0.390	0.215	**0.007** (−96.7%)	0.014 (−93.5%)	0.217 (+00.9%)
SemEnt	0.376	0.222	**0.008** (−96.4%)	0.010 (−95.5%)	0.247 (+11.3%)

Table 2: Comparison of ECE over accuracy of LLaVA-Med for Slake with the confidence reported by baseline ($Conf_{baseline}$), calibrated baseline ($Conf^{(1/T)}_{baseline}$), and the proposed calibration with grounding 2. Percentage improvement/regression in ECE via grounding from the calibrated baseline is also reported (in green/red) for each grounding model (GM).

its confidence score, and *Top-K* prompting, where the model produces K candidate responses with associated confidence scores in a single prompt, and the response with the highest confidence is selected. In contrast, two-stage approaches separate prediction from confidence estimation. *Self-reflection* prompting first generates an answer and then, in a subsequent pass, asks the model to reflect on that answer to produce a confidence estimate. *Visual Confidence-Aware Prompting (VCAP)* (Xuan et al., 2025) similarly follows a two-stage design, first eliciting a visual description along with a visual confidence score, and then performing the task conditioned on this self-assessed visual understanding to improve calibration.

5.5 Results

We report the average *Expected Calibration Error* (ECE) over accuracy with the confidence reported by (a) self-consistency baselines, (b) calibrated version of these baselines with temperature scaling, and (c) proposed calibration (2) for these baselines via grounding with different grounding models. Average ECE is calculated from 5 runs of random splits of the test/validation sets. Tables 1, and 2 show results for VQA and Slake, respectively. We observe a very low variance in all cases, and those values are in the Appendix. Values of hyperparameters T, C are also reported in the Appendix.

We also plot *reliability diagrams* for these baselines along with their temperature-scaled version, and the proposed grounding approach for calibration of these baselines with different grounding models (GM). These plots along with their respective ECE for LexSim, and SemEnt are reported in Fig. 3 for VQA and, Fig 8 in Appendix for Slake: LexSim and Sement are the top two UQ baselines in terms of ECE for both the datasets. Plots for other baselines (PredEnt and NumSets) are in the Appendix.

Observations on ECE: We make the following observations from Tables 1 and 2. First, ECE by the calibrated confidence via temperature scaling of all baselines is much lower than their original versions reported in the literature. This illustrates the efficacy of temperature scaling in calibrating the existing UQ techniques. Second, compared to confidence by the self-consistency approaches ($Conf_{baseline}$), the proposed grounding-based approach achieves much lower ECE in all but one test case for all grounding models on both VQA and Slake datasets, illustrating that the proposed approach is agnostic to the choice of GM. Finally, significant percentage improvements in ECE in most (all but three as shown in red) test cases - at least 32.5% for VQA and 34.3% for for Slake - from

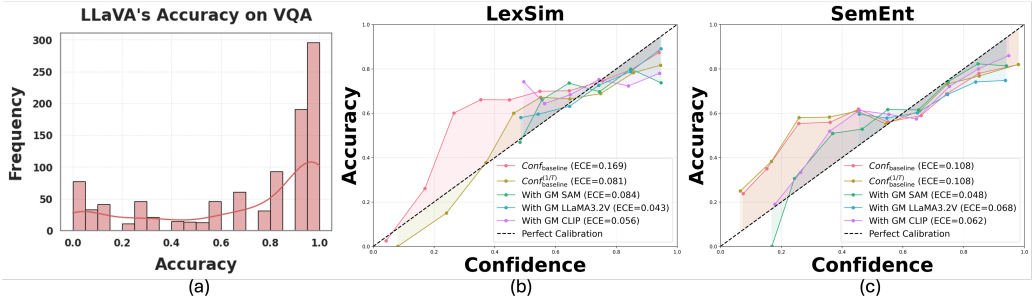


Figure 3: (a) Histogram on the frequency of LLaVA's accuracy on VQA. Reliability Diagrams for UQ of LLaVA on VQA by (b) LexSim, and (c) SemEnt baseline. Each diagram shows plots and the respective ECE for the confidence reported by the baseline $Conf_{baseline}$, its calibrated version $Conf_{baseline}^{(1/T)}$, and the proposed approach 2 for calibration with different GM.

the calibrated version of the baselines (as shown in green) illustrate that calibrating confidence of self-consistency scores on LLM's responses with external grounding is much more effective than calibrating those scores via temperature scaling.

Another important observation is that the amount of ECE improvement depends on the choice of the GM. In the case of VQA, LLaMA3.2V performs the best. Our hypothesis on this is as follows. LLaMA3.2V's diverse pretraining corpus has enhanced the model's capacity for commonsense reasoning and nuanced interpretation of complex image-text relationships, making it apt for grounding the responses of the VQA dataset that requires commonsense knowledge and understanding of visual domain. In the case of Slake, Biomed-Clip which is fine-tuned on medical domain performs better than the general-purpose LLaMA3.2V model. Although, Biomed-QwenVL is also fine-tuned for medical purposes but the use of synthetic data in the post-training might indicate its poor performance in the three baselines.

Observations from Reliability Diagrams: For VQA, LLaVA is able to answer most of the questions accurately - this is evident from the histogram on the frequency of LLaVA's accuracy on VQA 3 (a). This justifies the concentration of the reliability plots in the higher accuracy-confidence range (> 0.5) with all GM in case of LexSim and with LLaMA3.2V (best GM) on SemEnt. For other GM (SAM & CLIP) on SemEnt, the plot is more calibrated – closer to $x = y$ axis in all regions – in comparison to both the baseline and the temperature-scaled baseline.

For Slake, LLaVA-Med is not able to answer most of the questions accurately - this is evident from the histogram on the frequency of LLaVA-Med's accuracy on Slake 8 (a). This justifies the concentration of the reliability plots in the lower accuracy-confidence range (< 0.5) BiomedClip (best GM) on LexSim. The peaky behavior of the calibrated baseline and LlaMA3.2V in LexSim is due to uncalibrated confidence predictions on a very small number (2 to 3) inputs from higher (> 0.5) confidence bins. Similarly in case of SemEnt, BiomedLCip and LLaMA3.2V yields more calibrated curve in the lower accuracy-confidence range (< 0.5) but again we observe peaky behavior here, again due to uncalibrated confidence predictions on 2 to 3 inputs from higher confidence bins.

Additional robustness evidence (with results in Appendix): We provide several ablation studies to test whether the calibration gains are specific to a particular UQ baseline, dataset, VLM, grounding model, or calibration metric. First, Table 20 shows that the proposed grounding-based calibration also improves non-consistency-based verbalized confidence baselines, including CoT, Top-K, Self-Reflection, and VCAP. Second, Tables 14 and 15 show similar improvements on GQA and Whoops, extending the evaluation beyond the main VQA/Slake settings to compositional visual reasoning and caption generation. Third, Figure 11, Figure 12, and Tables 12–13 show that the approach remains effective across VLMs differing in openness and scale. We also report computation overhead and GM-only calibration baselines in Tables 5 and 6–11. Finally, in addition to ECE, Tables 16–19 report two additional calibration metrics, Smooth ECE (Blasiok and Nakkiran, 2024) and Brier score (Brier, 1950), to further evaluate the quality of the estimated confidence.

6 Conclusion

This work exposes a key limitation of self-consistency-based UQ for VLMs: **consistency**

does not imply accuracy. Consequently, relying solely on self-consistency can produce miscalibrated confidence estimates. We propose a calibrated UQ framework that complements self-consistency with cross-modal response consistency. Experiments across multiple VLMs, datasets, domains, tasks, and UQ baselines demonstrate that **cross-modal consistency, alongside self-consistency, is essential for achieving calibrated uncertainty estimates.**

Limitations

Our approach requires access to a grounding model (GM) capable of verifying model outputs against the input image, and introduces additional inference-time cost from this extra forward pass compared to self-consistency-only baselines. While our results show the method is robust to the specific choice of GM across the categories evaluated, the general framework assumes that a response can be meaningfully grounded back into the image (e.g., via bounding boxes or segmentation masks); it is less directly applicable to responses that are not spatially groundable, such as purely abstract reasoning steps. Separately, we observe that non-self-consistency verbalized-confidence baselines relying on structured output formats (e.g., XML/JSON) can fail to parse for vision-language models that do not reliably follow formatting instructions, which limits their comparability across all VLMs in our evaluation.

Acknowledgements

This material is based on work supported by the United States Air Force and Defense Advanced Research Projects Agency (DARPA) under Contract No. FA8750-23-C-0519, the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR0011-24-9-0424, and the Advanced Research Projects Agency for Health (ARPA-H) under Contract No. SP4701-23-C-0073. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the United States Air Force, Department of Defense, Defense Advanced Research Projects Agency (DARPA), Advanced Research Projects Agency for Health (ARPA-H) or the United States Government.

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement No. W911NF-17-2-0196. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Manoj Acharya, Karan Jariwala, and Christopher Kanan. 2019. Vqd: Visual query detection in natural scenes. *arXiv preprint arXiv:1904.02794*.

Amazon Web Services. 2023. Amazon EC2 P4 Instances. Accessed: August 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. 2014. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2616–2627.

Jaroslaw Blasiok and Preetum Nakkiran. 2024. Smooth ece: Principled reliability diagrams via kernel smoothing. In *International Conference on Learning Representations*, volume 2024, pages 1679–1707.

GW Brier. 1950. Verification of forecasts expressed in terms of probability,,,monthly weather review", 78 (1).

Daixuan Cheng, Shaohan Huang, Ziyu Zhu, Xintong Zhang, Wayne Xin Zhao, Zhongzhi Luan, Bo Dai, and Zhenliang Zhang. 2024. On domain-specific post-training for multimodal large language models. *arXiv preprint arXiv:2411.19930*.

Pranjal Chitale, Jay Gala, and Raj Dabre. 2024. An empirical study of in-context learning in llms for machine translation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7384–7406.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Morris H DeGroot and Stephen E Fienberg. 1983. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Loris Giulivi and Giacomo Boracchi. 2024. Explaining multi-modal large language models by analyzing their vision perception. *arXiv preprint arXiv:2405.14612*.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: A comprehensive evaluation of llms on creative writing. *arXiv preprint arXiv:2310.08433*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Tobias Groot and Matias Valdenegro-Toro. 2024. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. *arXiv preprint arXiv:2405.02917*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and

open questions. *ACM transactions on information systems*, 43(2):1–55.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Edwin T Jaynes. 1957. Information theory and statistical mechanics. *Physical review*, 106(4):620.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Hao Jiang, Qi Liu, Rui Li, Shengyu Ye, and Shijin Wang. 2024. Cursorcore: Assist programming through aligning anything. *arXiv preprint arXiv: 2410.07002*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Ramneet Kaur, Colin Samplawski, Adam D Cobb, Anirban Roy, Brian Matejek, Manoj Acharya, Daniel Elenius, Alexander Michael Berenbeim, John A Pavlik, Nathaniel D Bastian, and 1 others. 2024. Addressing uncertainty in llms to enhance reliability in generative ai. In *Neurips Safe Generative AI Workshop 2024*.

Vasily Kostumov, Bulat Nutfullin, Oleg Pilipenko, and Eugene Ilyushin. 2024. Uncertainty-aware evaluation for vision-language models. *arXiv preprint arXiv:2402.14418*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023b. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

Yukun Li, Sijia Wang, Lifu Huang, and Li-Ping Liu. 2024. Graph-based confidence calibration for large language models. *arXiv preprint arXiv:2411.02454*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, and 1 others. 2024b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8024–8035.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. 2023. Conformal language modeling. *arXiv preprint arXiv:2306.10193*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, and 1 others. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

Pritish Sahu, Karan Sikka, and Ajay Divakaran. 2024. Pelican: Correcting hallucination in vision-llms via claim decomposition and program of thought verification. *arXiv preprint arXiv:2407.02352*.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. 2024. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *arXiv preprint arXiv:2412.05563*.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236, Miami, Florida, US. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Sean Wang, Yicheng Jiang, Yuxin Tang, Lu Cheng, and Hanjie Chen. 2025. Copu: Conformal prediction for uncertainty quantification in natural language generation. *arXiv preprint arXiv:2502.12601*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations*.

Weihao Xuan, Qingcheng Zeng, Heli Qi, Junjue Wang, and Naoto Yokoya. 2025. Seeing is believing, but how much? a comprehensive analysis of verbalized calibration in vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1408–1450, Suzhou, China. Association for Computational Linguistics.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. *arXiv preprint arXiv:2401.12794*.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315.

Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. 2024. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, and 1 others. 2023a. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, and 1 others. 2023b. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, and 1 others. 2024. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature methods*, pages 1–11.

A Miscalibration of Existing SOTA self-consistency-based UQ approaches on TriviaQA and COQA

Fig. 4a and Fig. 4b show reliability diagrams for two self-consistency-based UQ approaches for predicting confidence of Llama-2-13B (Touvron et al., 2023) on TriviaQA (Joshi et al., 2017) and COQA (Reddy et al., 2019) datasets respectively; a common test setting considered by these approaches (Lin et al., 2023; Kuhn et al., 2023; Kaur et al., 2024). High expected calibration error (ECE) shows poor calibration of the model's accuracy with its confidence predicted by these approaches

B Prompt for the Foundation Grounding Models

Prompt

```
"I have an image and a short statement that describes a specific part of the image. Your job is to verify if this statement accurately reflects what is shown in the image.
Image: <Attached above>
Statement: "{response_text}"
Instructions: Respond with only one word — either "Yes" if the statement is correct, "No" if the statement is incorrect, or "Not sure" if you are uncertain. Do not provide any additional explanations."
```

We map "Yes" to the confidence score of 1, "No" as well as "Not sure" to the confidence score of 0.

C Examples of Question-Answer Pairs from VQAv2, GQA and Whoops Dataset

Fig. 5, and 6 shows sample multimodal question-answer pairs from the VQAv2 (Goyal et al., 2017) and GQA (Hudson and Manning, 2019) datasets, respectively. Fig. 7 shows some examples of the caption generation task for the Whoops dataset (Bitton-Guetta et al., 2023).

D Reliability Diagrams

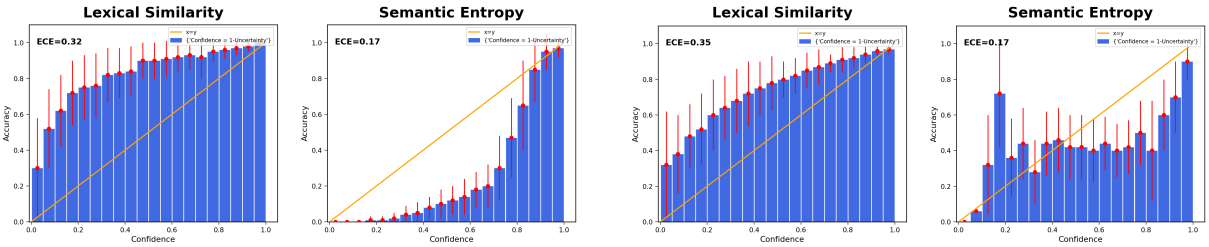
Figure 8 shows the reliability diagrams for LexSim and SemEnt on Slake. For Slake, LLaVA-Med is not able to answer most of the questions accurately, as shown by the accuracy histogram in Figure 8(a). This explains why the reliability plots are concentrated in the lower accuracy-confidence range. Figures 9 and 10 additionally show reliability diagrams for PredEnt and NumSets on VQA and Slake, respectively.

E ECE Results across different runs & Hyperparameter Configurations

Tables 3 show the values for mean and variance of ECE over 5 runs, as well as mean hyperparameter values of T and C again over 5 runs for VQA. Table 4 shows these values for Slake.

Baseline	Grounding Model	Mean ECE	Variance ECE	Mean T	Mean C
LexSim	NA	0.1692	0.00001	NA	NA
LexSim	Temp. Scaled Baseline	0.0813	0.00001	1.7	NA
LexSim	GroundedSAM	0.0524	0.00001	0.7	0.5
LexSim	Qwen-VL	0.0606	0.00001	1.1	0.5
LexSim	LLaMA 3.2V	0.0449	0.00001	0.3	0.5
LexSim	CLIP	0.0453	0.00001	0.7	0.5
NumSets	NA	0.4096	0.00001	NA	NA
NumSets	Temp. Scaled Baseline	0.2465	0.00001	5.1	NA
NumSets	GroundedSAM	0.1432	0.0001	0.3	0.5
NumSets	Qwen-VL	0.1249	0.0001	0.5	0.5
NumSets	LLaMA 3.2V	0.1265	0.00001	0.3	0.5
NumSets	CLIP	0.1434	0.00001	0.5	0.5
SemEnt	NA	0.1080	0.00001	NA	NA
SemEnt	Temp. Scaled Baseline	0.1078	0.00001	0.9	NA
SemEnt	GroundedSAM	0.0355	0.00001	2.9	0.2
SemEnt	Qwen-VL	0.0728	0.00001	0.9	0.5
SemEnt	LLaMA 3.2V	0.0290	0.0001	0.3	0.4
SemEnt	CLIP	0.0383	0.0001	4.7	0.1
PredEnt	NA	0.4445	0.0001	NA	NA
PredEnt	Temp. Scaled Baseline	0.1904	0.0002	9.7	NA
PredEnt	GroundedSAM	0.1145	0.0001	0.7	0.5
PredEnt	Qwen-VL	0.1187	0.0001	2.7	0.5
PredEnt	LLaMA 3.2V	0.0964	0.0001	0.3	0.5
PredEnt	CLIP	0.1165	0.0001	0.9	0.5

Table 3: Mean Expected Calibration Error (ECE), variance of ECE across five random splits over test/validation sets across different grounding models and baselines for the VQA dataset, and the corresponding average values of the temperature scaling parameter (T) and confidence threshold (C). Results are presented for all the four baselines—LexSim, NumSets, SemEnt, and PredEnt—each evaluated with the vanilla baseline, temperature-scaled baseline, and multiple vision-language grounding models including GroundedSAM, Qwen-VL, LLaMA 3.2V, and CLIP.



(a) TriviaQA (b) COQA

Figure 4: Reliability diagrams with the expected accuracy of LLaMA-2-13B on TriviaQA and COQA plotted as a function of the model's confidence predicted by two SOTA self-consistency-based UQ approaches. A perfect calibration between the model's accuracy and the predicted confidence would result in red points, representing the average accuracy for each confidence bin, lying on the $x = y$ axis with low variance.

F Verbalized Prompt Templates & Limitations

We provide the complete prompt templates used for each verbalized uncertainty quantification method.

F.1 Chain-of-Thought (CoT)

The CoT prompt elicits step-by-step reasoning followed by a structured answer and a verbalized confidence score in a single forward pass.

CoT

```
USER: <image>
Task: Answer the following question
    about the image. Provide your
    best answer along with a
    confidence score (0% to 100%).

Instructions:
- Carefully read and analyze the
    question.
- Reason through the answer step by
    step.
- At the end, present your final
    answer and a confidence score in
    the following XML format:
<answer>your answer here</answer>
<confidence>your confidence score
    here</confidence>

Example:
Question: How many people are
    visible in the image?

A: Let me analyze the image
    carefully. I can see several
    individuals in the scene. On the
    left side, there are two people
    standing together. In the
    center, there is one person
    sitting. On the right side, I
    can see one more person walking.
    Counting all visible
    individuals: 2 + 1 + 1 = 4
    people total.
<answer>4</answer>
<confidence>85%</confidence>

Now, here is your question:
Question: {question}

ASSISTANT:
```

F.2 Top-K Prompting

The Top-K prompt (with K=3) asks the model to produce K ranked answer candidates, each with an associated confidence. If fewer than K distinct answer options are plausible, the model is instructed to score only the available options.

Top-K Prompting, ($K = 3$)

```
USER: <image>
You are given an image and a visual
    question.

YOUR TASK is to answer the visual
    question.

Instructions:
- Provide your 3 best guesses and
    the probability that each is
    correct (0% to 100%).
- If fewer than 3 options exist,
    only provide the available
    answers with their confidence
    scores.
- Please reason step by step.
- At the end, present your final
    answers in the following
  XML format:
<answer1>[your first best
    answer]</answer1>
<confidence1>[confidence for
    answer1]</confidence1>
<answer2>[your second best
    answer]</answer2>
<confidence2>[confidence for
    answer2]</confidence2>
<answer3>[your third best
    answer]</answer3>
```



Q1: What does the sign say?
A: No cursing



Q2: What color is the engine?
A: Red



Q3: Are most of the people wearing hats?
A: No

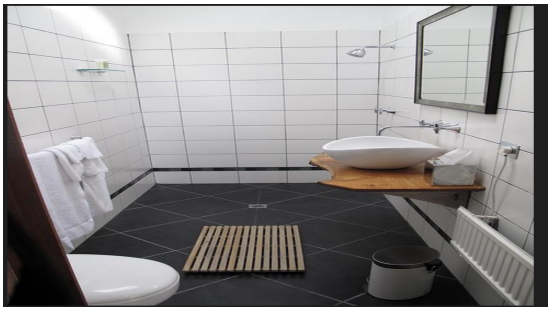


Q4: Who does the man on the right resemble?
A: Surfer

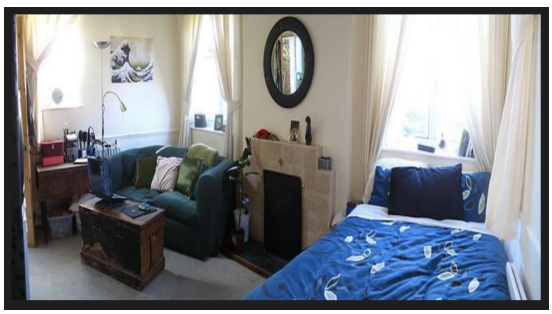
Figure 5: Example Q&A pairs from the VQAv2 dataset (Goyal et al., 2017).



Q1: Is the blue bag to the right or to the left of the player that is to the left of the towel?
A: The bag is to the left of the player.



Q2: Is the shower near the mirror dirty and large?
A: No, the shower is large but clean.



Q3: Is the fireplace near the couch white and closed?
A: No, the fireplace is closed but black.



Q4: Does the umbrella look open and large?
A: No, the umbrella is open but small.

Figure 6: Example Q&A pairs from the GQA dataset (Hudson and Manning, 2019).



Q1: Describe the image.
A: A Renaissance woman with a cheeseburger.



Q2: Describe the image.
A: A pagoda sits in front of the Eiffel Tower.



Q3: Describe the image.
A: A man is hiking up a mountain wearing a suit.



Q4: Describe the image.
A: Vikings ride on public transportation.

Figure 7: Example Image-caption pairs from the Whoops dataset (Bitton-Guetta et al., 2023).

Baseline	Grounding Model	Mean ECE	Variance ECE	Mean T	Mean C
LexSim	NA	0.0308	0.00001	NA	NA
LexSim	Temp. Scaled Baseline	0.0312	0.00001	0.9	NA
LexSim	BiomedCLIP	0.0040	0.00001	0.1	0.2
LexSim	LLaMA 3.2V	0.0134	0.00001	0.5	0.1
NumSets	NA	0.4263	0.00001	NA	NA
NumSets	Temp. Scaled Baseline	0.2015	0.00001	0.1	NA
NumSets	BiomedCLIP	0.0070	0.00001	0.1	0.2
NumSets	LLaMA 3.2V	0.0212	0.00001	0.1	0.2
SemEnt	NA	0.3764	0.00001	NA	NA
SemEnt	Temp. Scaled Baseline	0.2221	0.00001	0.1	NA
SemEnt	BiomedCLIP	0.0080	0.00001	0.1	0.2
SemEnt	LLaMA 3.2V	0.0098	0.00001	0.3	0.2
PredEnt	NA	0.3902	0.00001	NA	NA
PredEnt	Temp. Scaled Baseline	0.2149	0.00001	0.1	NA
PredEnt	BiomedCLIP	0.0069	0.00001	0.1	0.2
PredEnt	LLaMA 3.2V	0.0138	0.00001	0.1	0.2

Table 4: Mean Expected Calibration Error (ECE), variance of ECE across five random splits over test/validation sets across different grounding models and baselines for the Slake dataset, and the corresponding average values of the temperature scaling parameter (T) and confidence threshold (C). Results are presented for all the four baselines—LexSim, NumSets, SemEnt, and PredEnt—each evaluated with the vanilla baseline, temperature-scaled baseline, and two grounding models: BiomedCLIP (a biomedical domain-specific model) and LLaMA 3.2V.

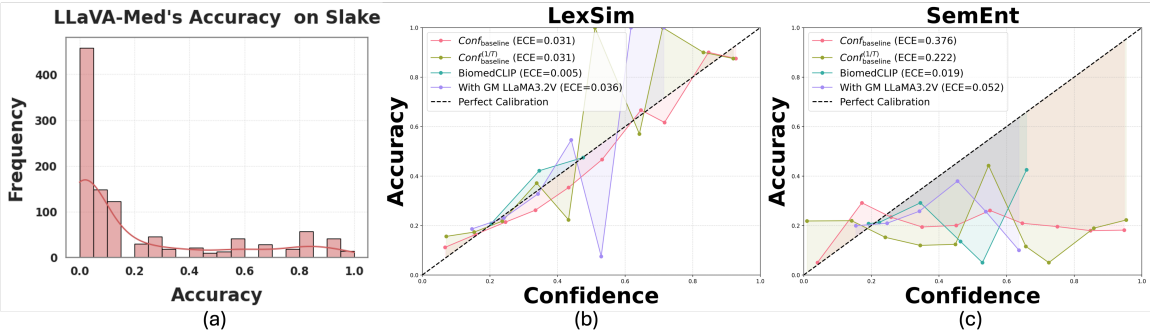


Figure 8: (a) Histogram on the frequency of LLaVA-Med's accuracy on Slake. Reliability diagrams for UQ of LLaVA-Med on Slake by (b) LexSim, and (c) SemEnt. Each diagram shows plots and the respective ECE for the confidence reported by the baseline $Conf_{\text{baseline}}$, its calibrated version $Conf_{\text{baseline}}^{(1/T)}$, and the proposed approach (2) for calibration with different grounding models.

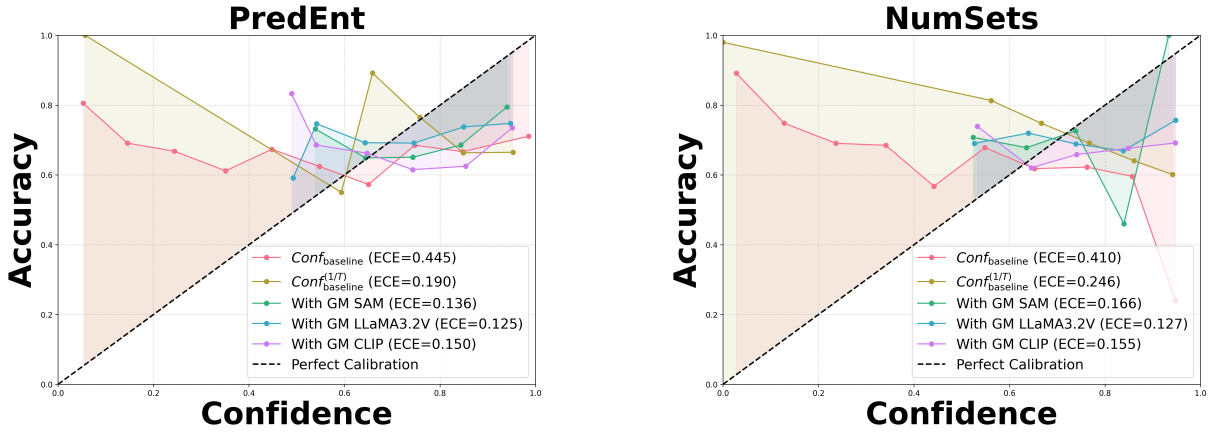


Figure 9: Reliability diagrams for UQ by PredEnt and NumSets baselines for LLaVA on the VQA dataset. Each diagram shows plots and ECE for the confidence reported by the baseline $Conf_{\text{baseline}}$, its calibrated version $Conf_{\text{baseline}}^{(1/T)}$, and the proposed approach (2) for calibration with different grounding models.

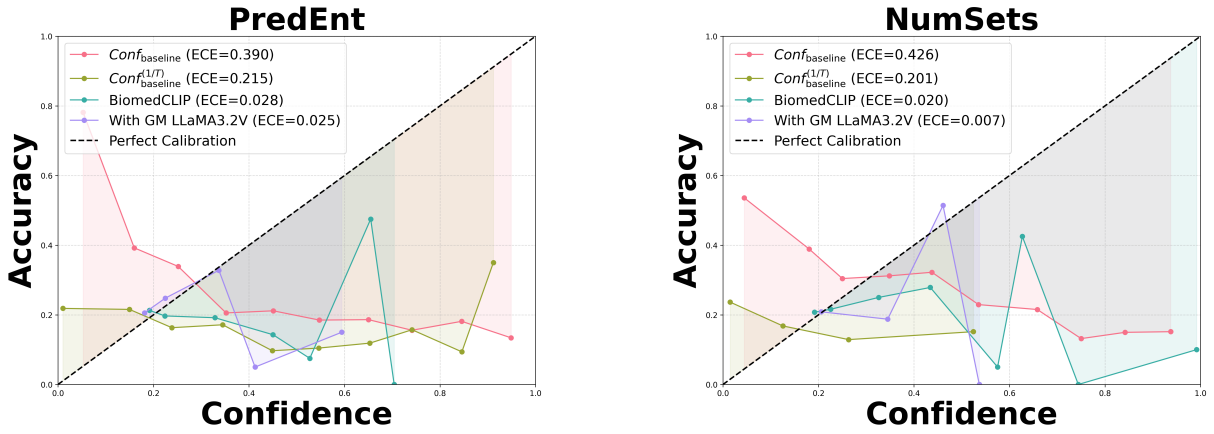


Figure 10: Reliability diagrams for UQ by PredEnt and NumSets baselines for LLaVA-Med on Slake. Each diagram shows plots and ECE for the confidence reported by the baseline $Conf_{\text{baseline}}$, its calibrated version $Conf_{\text{baseline}}^{(1/T)}$, and the proposed approach (2) for calibration with different grounding models.

```
<confidence3>[confidence for
    answer3]</confidence3>
```

```
Example output:
[YOUR_REASONING]
<answer1>yes</answer1>
<confidence1>75%</confidence1>
<answer2>no</answer2>
<confidence2>20%</confidence2>
<answer3>maybe</answer3>
<confidence3>5%</confidence3>

Question: {question}
ASSISTANT:
```

F.3 Self-Reflection

Self-Reflection is a two-stage method. In **Round 1**, the model answers the question with step-by-step reasoning but *does not* output a confidence score. In **Round 2**, the model receives the original question together with its Round 1 answer and is asked to reflect and output a single verbalized confidence score.

Self-Reflection — Round 1

```
USER: <image>
You are given an image and a visual
    question.

YOUR TASK is to answer the visual
    question.

Instructions:
- Please reason step by step.
- At the end, present your final
    answer in the following
  XML format:
<answer>[your answer to the
    question]</answer>

Example output:
[YOUR_REASONING]
<answer>yes</answer>

Question: {question}
ASSISTANT:
```

Self-Reflection — Round 2

```
USER: <image>
Task: Reflect on the following
    problem and solution, and
    provide a final confidence score.

Instructions:
- Carefully read and analyze the
    problem and solution.
- Reason through the solution step
    by step.
- At the end, present your final
    confidence score in the
    following XML format:
<confidence>confidence score
    here</confidence>

Example output:
[YOUR_REASONING]
<confidence>80%</confidence>

Problem: {question}
Solution: {solution}
ASSISTANT:
```

F.4 VCAP (Visual Confidence-Aware Prompting)

VCAP is a two-stage method. **Round 1** prompts the model to produce a detailed image description and attach a confidence score to that description. **Round 2** injects the description (including its confidence) and asks the model to answer the target question with confidence.

VCAP — Round 1: Image Description

```
USER: <image>
Task: Describe the image with enough
    detail to help answer visual
    questions about it. Provide a
    confidence score
(0% to 100%).

Instructions:
- Observe all important elements in
    the image.
- Describe objects, people, actions,
    colors, and spatial
    relationships.
- Provide a confidence score
    indicating certainty about your
    description.
- Please reason step by step.
- Present your output in XML format:
<description>your detailed
    description</description>
<confidence>your confidence
    score</confidence>

Example:
ASSISTANT: Looking at this image: In
    the foreground, I see a wooden
    table with a red tablecloth. On
    the table: a white coffee mug on
    the left, a plate with food in
    the center, and silverware on
    the right. The background shows
    a kitchen with white cabinets
    and a window.
    <description>Wooden table with a
    red tablecloth, a white coffee
    mug, a plate with food, and
    silverware. Background: kitchen
    with white cabinets and a
    window.</description>
<confidence>90%</confidence>
```

```
Now, please describe this image:
ASSISTANT:
```

VCAP — Round 2: Question Answering

```
USER: <image>
Task: Answer the following question
    about the image. Use both the
    image and the description below.

You have previously generated this
    description:
{description_with_confidence}

Instructions:
- Reason step by step, considering
    both the image and your own
    description.
- Take into account the confidence
    score of the description.
- At the end, present your final
    answer and a confidence score
    (0% to 100%) in XML format:
<answer>your answer here</answer>
<confidence>your confidence score
    here</confidence>

Example:
Previous description: "A park with
    green grass, several people
    playing with a frisbee, trees in
    the background,
sunny day" (confidence: 85%)

Question: What activity is happening?

A: Based on my description
    mentioning people playing with a
    frisbee (85% confidence), and
    checking the image to confirm, I
    see the frisbee in mid-air and
    people in throwing/catching
    positions.
<answer>playing frisbee</answer>
<confidence>90%</confidence>

Question: {question}

A:
```

A practical limitation of verbalized confidence estimation is its reliance on structured output formats (e.g., XML/JSON) for reliable parsing, which introduces two failure modes. First, many vision-language models—especially domain-specialized variants—do not consistently adhere to strict formatting constraints, resulting in parsing failures. This failure was evident for LLaVA-Med in our experiments, where the model could not generate a self-confidence score for these baselines. Second, imposing structured output constraints can itself degrade task performance; Tam et al. (2024) report drops of up to 38 percentage points on mathematical reasoning benchmarks under structured-output requirements (Tam et al., 2024).

G Additional Ablations

1. Computation Cost Analysis: Table 5 reports the time taken by baselines and grounding models on Slake. We acknowledge that our calibration approach introduces computational overhead due to the use of an external model for grounding. However, as evident from Tables 1 and 2, we would like to emphasize that we get substantial performance gains even with a lightweight GM such as Clip with at most 0.01% computational overhead.

Baseline	Grounding Model
LexSim: 44.24	BiomedClip: 0.47
NumSets: 47.82	Llama3.2V: 9.44
PredEnt: 44.18	Biomed-QwenVL: 17.68
SemEnt: 47.82	

Table 5: Computation Cost Analysis (in Seconds) on Slake

2. Using $Conf_{GM}^{1/T}$ as a Baseline: We also report ECE with temperature-scaled grounding model confidence as another baseline. These results are shown in Tables 6, 7, 8 show results for $Conf_{GM}^{(1/T)}$ as the baseline for VQA with Clip, LLaMA3.2V, and QwenVL as the grounding model respectively. Tables 9, 10, 11 show results for $Conf_{GM}^{(1/T)}$ as the baseline for Slake with Biomed-Clip, LLaMA3.2V, and QwenVL as the grounding model respectively. The proposed approach outperforms this baseline on all but two 2 cases on VQA and all but 3 test cases on Slake.

3. Different Sizes of VLMs For ablations on different sizes of VLMs, in addition to Llava-7B, we perform experiments with Phi-4-multimodal-instruct-5.6B (Abouelenin et al., 2025) and Qwen2.5-VL-32B-Instruct (Bai et al., 2025) models on GQA with Clip as the grounding model. Figure 11 shows significant percentage improvements – ranging from 41.6% to 97.7% – by our approach in ECE from the temperature-scaled baselines for all models. Tables 12 and 13 show detailed results for Phi and Qwen, respectively.

4. Generalizability w.r.t Datasets and Tasks: Tables 14 and 15, show these results for all the baselines with Llava-v1.5-7B as the VLM and Clip as the GM for Whoops and GQA respectively. Here,

Baseline	Without Grounding (↓)		With Grounding (↓)	
	$Conf_{\text{baseline}}$	$Conf_{\text{baseline}}^{(1/T)}$	$Conf_{GM}^{(1/T)}$	$Conf_{\text{baseline}} \times Conf_{GM}^{(1/T)} + C$
LexSim	0.169	0.081	0.11 (+35.8%)	**0.045** (–44.6%)
NumSets	0.410	0.246	**0.11** (–55.3%)	0.143 (–41.8%)
PredEnt	0.445	0.190	**0.11** (–42.1%)	0.117 (–38.3%)
SemEnt	0.108	0.108	0.11 (+1.90%)	**0.038** (–64.6%)

Table 6: Temperature-Scaled Grounding as baseline with Clip as the GM on VQA.

Baseline	Without Grounding (↓)		With Grounding (↓)	
	$Conf_{\text{baseline}}$	$Conf_{\text{baseline}}^{(1/T)}$	$Conf_{GM}^{(1/T)}$	$Conf_{\text{baseline}} \times Conf_{GM}^{(1/T)} + C$
LexSim	0.169	0.081	0.14 (+72.8%)	**0.045** (–44.6%)
NumSets	0.410	0.246	0.14 (–43.1%)	**0.127** (–48.4%)
PredEnt	0.445	0.190	0.14 (–26.3%)	**0.096** (–49.4%)
SemEnt	0.108	0.108	0.14 (+29.6%)	**0.029** (–73.1%)

Table 7: Temperature-Scaled Grounding as baseline with LLaMA3.2V as the GM on VQA.

Baseline	Without Grounding (↓)		With Grounding (↓)	
	$Conf_{\text{baseline}}$	$Conf_{\text{baseline}}^{(1/T)}$	$Conf_{GM}^{(1/T)}$	$Conf_{\text{baseline}} \times Conf_{GM}^{(1/T)} + C$
LexSim	0.169	0.081	0.31 (+282.7%)	**0.061** (–25.0%)
NumSets	0.410	0.246	0.31 (+26.00%)	**0.125** (–49.2%)
PredEnt	0.445	0.190	0.31 (+63.20%)	**0.119** (–37.5%)
SemEnt	0.108	0.108	0.31 (+187.0%)	**0.073** (–32.5%)

Table 8: Temperature-Scaled Grounding as baseline with QwenVL as the GM on VQA.

Baseline	Without Grounding (↓)		With Grounding (↓)	
	$Conf_{\text{baseline}}$	$Conf_{\text{baseline}}^{(1/T)}$	$Conf_{\text{GM}}^{(1/T)}$	$Conf_{\text{baseline}} \times Conf_{GM}^{(1/T)} + C$
LexSim	0.031	0.031	0.14 (+351.6%)	**0.004** (–87.1%)
NumSets	0.426	0.201	0.14 (–30.30%)	**0.007** (–96.5%)
PredEnt	0.390	0.215	0.14 (–34.90%)	**0.007** (–96.7%)
SemEnt	0.376	0.222	0.14 (–36.90%)	**0.008** (–96.4%)

Table 9: Temperature-Scaled Grounding as baseline with BiomedClip as the GM on Slake.

Baseline	Without Grounding (↓)		With Grounding (↓)	
	$Conf_{\text{baseline}}$	$Conf_{\text{baseline}}^{(1/T)}$	$Conf_{GM}^{(1/T)}$	$Conf_{\text{baseline}} \times Conf_{GM}^{(1/T)} + C$
LexSim	0.031	0.031	0.17 (+448.4%)	**0.013** (–58.1%)
NumSets	0.426	0.201	0.17 (–15.40%)	**0.021** (–89.6%)
PredEnt	0.390	0.215	0.17 (–20.90%)	**0.014** (–93.5%)
SemEnt	0.376	0.222	0.17 (–23.40%)	**0.010** (–95.5%)

Table 10: Temperature-Scaled Grounding as baseline with LLaMA3.2V as the GM on Slake.

Baseline	Without Grounding (↓)		With Grounding (↓)	
	$Conf_{\text{baseline}}$	$Conf^{(1/T)}_{\text{baseline}}$	$Conf^{(1/T)}_{GM}$	$Conf_{\text{baseline}} \times Conf^{(1/T)}_{GM} + C$
LexSim	**0.031**	**0.031**	0.10 (+222.6%)	0.038 (+22.6%)
NumSets	0.426	0.201	**0.10** (–50.20%)	0.132 (–34.3%)
PredEnt	0.390	0.215	**0.10** (–53.50%)	0.217 (+0.9%)
SemEnt	0.376	0.222	**0.10** (–54.90%)	0.247 (+11.3%)

Table 11: Temperature-Scaled Grounding as baseline with QwenVL as the GM on Slake.

Baseline	Without Grounding (Ours) (↓)		With Grounding (Ours) (↓)
	$Conf_{\text{baseline}}$	$Conf^{(1/T)}_{\text{baseline}}$	**CLIP**
LexSim	0.136	0.095	**0.024** (–74.7%)
NumSets	0.350	0.318	**0.016** (–95.0%)
SemEnt	0.194	0.139	**0.022** (–84.2%)
PredEnt	0.375	0.287	**0.017** (–94.1%)

Table 12: Baseline comparisons of ECE for **Phi** on Q&A task of GQA. Percentage improvement in ECE via grounding is from the calibrated baseline $Conf^{(1/T)}_{baseline}$.

Baseline	Without Grounding (↓)		With Grounding (Ours) (↓)
	$Conf_{\text{baseline}}$	$Conf^{(1/T)}_{\text{baseline}}$	**CLIP**
LexSim	0.157	0.064	**0.002** (–96.9%)
NumSets	0.240	0.174	**0.005** (–97.1%)
SemEnt	0.387	0.388	**0.009** (–97.7%)
PredEnt	0.375	0.305	**0.032** (–89.6%)

Table 13: Baseline comparisons of ECE for **Qwen** on Q&A task of GQA. Percentage improvement in ECE via grounding is from the calibrated baseline $Conf^{(1/T)}_{baseline}$.

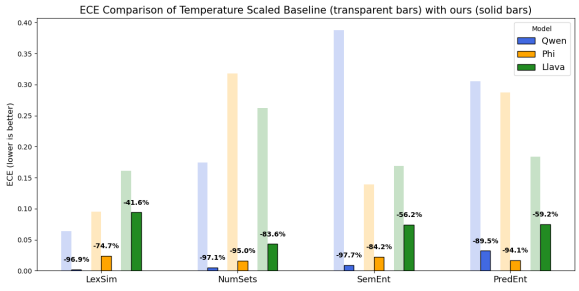


Figure 11: Temperature-scaled baselines' comparisons of ECE (↓) with Clip as the GM in the proposed approach for different VLMs – Qwen, Phi, Llava – on GQA.

Baseline	Without Grounding (↓)		With Grounding (Ours) (↓)
	$Conf_{\text{baseline}}$ *uncalibrated*	$Conf^{(1/T)}_{\text{baseline}}$ *calibrated*	**CLIP** *ours*
LexSim	0.192	0.161	**0.094** (–41.6%)
NumSets	0.318	0.262	**0.043** (–83.6%)
SemEnt	0.214	0.169	**0.074** (–56.2%)
PredEnt	0.339	0.184	**0.075** (–59.2%)

Table 14: Baseline comparisons of ECE for **LLaVA** on Q&A task of GQA. Percentage improvement in ECE via grounding is from the calibrated baseline $Conf^{(1/T)}_{baseline}$.

also we observe significant improvements in the ECE for both GQA (at least 41.6%) and Whoops (at least 68.6%).

H Computation Resources

All experiments were conducted on a combination of AWS EC2 `p4de.24xlarge` and `p4d.24xlarge` instances. The `p4de.24xlarge` configuration includes 8 × NVIDIA A100 GPUs (80 GB HBM2e per GPU, totaling 640 GB of GPU memory), 96 vCPUs, 1,152 GiB of system RAM, 8 TB of local NVMe SSD storage, 400 Gbps network bandwidth, and approximately 19 Gbps of EBS throughput. The `p4d.24xlarge` instances, in comparison, provide 8 × NVIDIA A100 GPUs (40 GB HBM2 per GPU), 1.1 TiB RAM, and similar networking and storage specifications (Amazon Web Services, 2023). Our codebase was implemented in Python using PyTorch (Paszke et al., 2019), with model development and inference supported by the Hugging Face Transformers library (Wolf et al., 2019).

I Additional Calibration Metrics

Tables 16–19 report Smooth ECE and Brier score on VQA for CLIP and LLaMA3.2-V across all four baselines.

Baseline	Without Grounding (↓)		With Grounding (↓)
	$Conf_{\text{baseline}}$	$Conf^{(1/T)}_{\text{baseline}}$	**CLIP**
LexSim	0.125	0.140	**0.044** (–68.6%)
NumSets	0.406	0.199	**0.060** (–69.8%)
SemEnt	0.484	0.352	**0.053** (–84.9%)
PredEnt	0.639	0.545	**0.030** (–94.5%)

Table 15: Baseline comparisons of ECE for LLaVA on caption generation task of Whoops. Percentage improvement in ECE via grounding is from $Conf^{(1/T)}_{baseline}$.

Baseline	T	C	Baseline Smooth ECE	Calibrated Smooth ECE
LexSim	3.3833	0.3860	0.1803	0.0877
NumSets	15.2752	0.5000	0.3760	0.2034
SemEnt	3.1030	0.2040	0.0994	0.0819
PredEnt	8.1881	0.4960	0.4038	0.1799

Table 16: Smooth ECE results on VQA with CLIP as the grounding model.

Baseline	T	C	Baseline Brier	Calibrated Brier
LexSim	3.3833	0.3860	0.1498	0.1240
NumSets	15.2752	0.5000	0.3626	0.1814
SemEnt	3.1030	0.2040	0.1196	0.1210
PredEnt	8.1881	0.4960	0.3709	0.1670

Table 17: Brier score results on VQA with CLIP as the grounding model.

Baseline	T	C	Baseline Smooth ECE	Calibrated Smooth ECE
LexSim	1.6215	0.4100	0.1803	0.0801
NumSets	7.6275	0.4980	0.3760	0.1661
SemEnt	4.3042	0.3280	0.0994	0.1208
PredEnt	0.8608	0.4960	0.4038	0.1514

Table 18: Smooth ECE results on VQA with LLaMA3.2-V as the grounding model.

Baseline	T	C	Baseline Brier	Calibrated Brier
LexSim	1.6215	0.4100	0.1498	0.1152
NumSets	7.6275	0.4980	0.3626	0.1562
SemEnt	4.3042	0.3280	0.1196	0.1333
PredEnt	0.8608	0.4960	0.3709	0.1499

Table 19: Brier score results on VQA with LLaMA3.2-V as the grounding model.

J Ablation Studies

1. Generalizability w.r.t the underlying UQ baseline. We conduct ablations on LLaVA for VQA for

all four non-consistency-based verbalized prompting UQ baselines. Table 20 shows these results. Here also, the proposed grounding-based calibration consistently outperforms all four verbalized prompting baselines across all grounding models, achieving improvements ranging from 60.8% to 96.4% in ECE as compared to temperature-scaled UQ baselines. Details on the verbalized prompts for these baselines and their limitations are listed in Appendix.

2. Generalizability w.r.t Datasets and Tasks: Apart from VQA for general Q&A and Slake for medical Q&A, we also consider two more datasets, namely GQA for Q&A and Whoops for the caption generation. These datasets are chosen as they test LLMs at varying levels of visuo-linguistic reasoning. Tables 14 and 15 in Appendix, show these results for all the baselines with Llava-v1.5-7B as the VLM and Clip as the GM. Here, also we observe significant improvements in the ECE for both GQA (at least 41.6%) and Whoops (at least 68.6%).

3. Generalizability w.r.t VLMs: Here we consider variability in terms of closed model such as Gemini-2.5-pro (Comanici et al., 2025), and different sizes of the VLMs. Unlike LLaVA, Gemini is a closed model without access to its generated token probabilities. As a result, baselines such as NumSets, PredEnt, and SemEnt, which rely on these probabilities, cannot be tested. So, we compare using the LexSim baseline, which assesses lexical similarity between every pair of responses. Figure 12 shows up to 86.25% improvements in ECE from the temperature-scaled baseline by the proposed approach with all grounding models on Slake. Similar results with ablations on different sizes of VLMs with parameters ranging from 5.6B (in Phi4) to 32B (in Qwen2.5) are included in the Appendix: Figure 11 and Tables 12 and 13.

We also perform other ablations on a) computation overhead due to the use of an external GM (Table 5) acknowledging that we get substantial performance gains even with a lightweight GM such as Clip with at most 0.01% computational overhead, and b) using temperature-scaled GM as baseline (Tables 6, 7, 8, 9, 10, 11). Details and results of these ablations are included in Appendix.

K Additional Calibration Metrics

To complement the standard ECE evaluation, we additionally report two calibration metrics: Smooth ECE (Blasiok and Nakkiran, 2024) and Brier score.

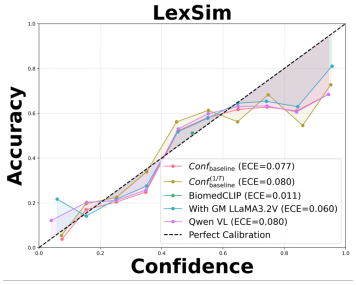


Figure 12: Reliability diagram and ECE comparison of LexSim with different GM in the proposed calibration approach for Gemini on the medical dataset Slake.

We evaluate these metrics on VQA across two GMs, CLIP and LLaMA3.2V, for all four baselines. Smooth ECE improvements range from 17.6% to 55.6%, and Brier score improvements range from 17.2% to 56.9%, with the exception of SemEnt, which shows marginal regression (-1.1% to -11.4%). These results provide additional evidence that the proposed approach improves confidence calibration beyond the standard ECE metric. Detailed results are reported in Tables 16–19.

Baseline	Without Grounding (↓)		With Grounding (↓)			
	$Conf_{\text{baseline}}$	$Conf^{(1/T)}_{\text{baseline}}$	GroundedSAM	CLIP	LLaMA3.2V	QwenVL
CoT	0.273	0.176	0.039 (-85.9%)	0.019 (-93.2%)	0.021 (-92.4%)	0.010 (-96.4%)
Top-K	0.297	0.126	0.010 (-92.4%)	0.036 (-71.4%)	0.049 (-60.8%)	0.017 (-86.5%)
Self-Reflection	0.410	0.334	0.056 (-83.2%)	0.011 (-96.6%)	0.103 (-69.2%)	0.133 (-60.3%)
VCAP	0.279	0.198	0.015 (-94.7%)	0.008 (-97.1%)	0.093 (-66.7%)	0.073 (-73.7%)

Table 20: Comparison of ECE over accuracy of LLaVA for VQA with confidence reported by non-consistency-based verbalized prompting UQ baselines ($Conf_{baseline}$), calibrated baseline ($Conf^{(1/T)}_{baseline}$), and our approach with grounding 2. Percentage improvement in ECE via grounding from the calibrated baseline is also reported (in green) for each GM.